Sammelband zur Textverarbeitung

FORSCHUNGSBERICHTE DES

INSTITUTS FÜR DEUTSCHE SPRACHE

MANNHEIM

herausgegeben von
Ulrich Engel und Gerhard Stickel

Schriftleitung: Eva Teubert

Band 3

Peter Kern

Bemerkungen zum Problem der Textklassifikation

Manfred W. Hellmann

Über Corpusgewinnung und Dokumentation im Mannheimer Institut für deutsche Sprache

Werner Müller

Teilerhebungen und ihre Anwendung auf die Sprachbearbeitung

Ulrich Engel

Das Mannheimer Corpus

TBL Verlag Gunter Narr · Tübingen

Tübingen 1976

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Kern, Peter

Bemerkungen zum Problem der Textklassifikation / Peter Kern. Über Corpusgewinnung und Dokumentation im Mannheimer Institut für Deutsche Sprache / Manfred W. Hellmann. Teilerhebungen und ihre Anwendung auf die Sprachbearbeitung / Werner Müller. Das Mannheimer Corpus / Ulrich Engel. Unveränd. Nachdr. d. 1. Aufl. 1968. — Tübingen: TBL-Verlag Narr, 1976.

(Forschungsberichte / Institut für Deutsche Sprache Mannheim; Bd. 3)

Nebent.: Sammelband zur Textverarbeitung.

ISBN 3—87808—603—2.

NE: Hellmann, Manfred W.: Über Corpusgewinnung und Dokumentation im Mannheimer Institut für Deutsche Sprache; Müller, Werner: Teilerhebungen und ihre Anwendung auf die Sprachbearbeitung; Engel, Ulrich: Das Mannheimer Corpus; NT.

unveränderter Nachdruck der 1. Auflage 1968

ISBN 3–87808–603–2

Vorbemerkungen

Der dritte Band der Forschungsberichte des Instituts für deutsche Sprache enthält verschiedene Beiträge zum Problem der Texte, die sprachwissenschaftlichen Untersuchungen zugrunde zu legen sind.

Peter Kern entwirft ein Verfahren zur Festlegung von Texttypen, Manfred Hellmann stellt die Frage nach der Repräsentativität eines Corpus, und Werner Müller zeigt Möglichkeiten der Anwendung statistischer Methoden in linguistischen Untersuchungen. Damit schließen sich die Betrachtungen, besonders die letzte, auch thematisch an Günther Billmeiers Studie "Über die Signifikanz von Auswahltexten" an, die in Band 2 der Forschungsberichte des Instituts für deutsche Sprache erschienen ist. Die hier genannten Arbeiten enthalten wichtige Vorschläge und Hinweise, daneben auch Kritik an manchen bisher geübten Verfahren. Dies war erwünscht und beabsichtigt. Der Frage der Corpusgewinnung war die gemeinsame Sitzung der Kommissionen für Dokumentation der deutschen Gegenwartssprache und für datenverarbeitende Maschinen und Sprachforschung am 29.6.1968 in Mannheim gewidmet, und auf dieser Sitzung trugen die vier Autoren ihre zum Teil voneinander abweichenden Thesen vor, die hier nochmals zur Diskussion gestellt werden. Schließlich legt Ulrich Engel in einem kurzen Beitrag dar, welche Erwägungen für die Zusammenstellung des Corpus in seiner heutigen Gestalt maßgebend waren.

Die Herausgeber

Inhalt

	Seite
Peter Kern: Bemerkungen zum Problem der Textklassifikation	3
Manfred W. Hellmann: Über Corpusgewinnung und Dokumentation im Mannheimer Institut für deutsche Sprache	25
Werner Müller: Teilerhebungen und ihre Anwendung auf die Sprachbearbeitung	55
Ulrich Engel: Das Mannheimer Corpus	75

Bemerkungen zum Problem der Textklassifikation

von Peter Kern

Übersicht:

1. Problematik und Lösungsversuche (3)

2. Versuch einer Typik der geschriebenen Sprache entsprechend der Schreibintention (8)

3. Anwendung (13)

Anmerkungen (22)

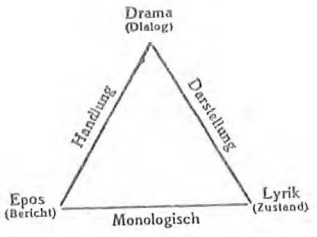
1. Problematik und Lösungsversuche

Der vorliegende Versuch sieht sich folgendem Problem gegenüber:
Er soll ein System erstellen, das so vollständig wie möglich - im Idealfall absolut lückenlos - alle denkbaren schriftlichen Manifestationen unserer Sprache repräsentiert, d.h. alle denkbaren Gesichtspunkte vereinigt, unter denen Texte entstehen, stehen und gelesen werden. Er soll andererseits auch eine praktikable Grundlage liefern für die Auswahl eines repräsentativen und dennoch überschaubaren Dokumentationskorpus.
Diesem Dilemma wird man nicht endgültig entgehen können angesichts der Fülle und Vielfalt des Materials. Man wird, wenn der Versuch überhaupt sinnvoll sein soll, eine Methode suchen müssen, die von eben dieser Fülle zunächst absieht und vielmehr versucht, von außen her, also deduktiv, nach Gesetzen zu forschen, die für Produktion und Konstitution von Texten maßgeblich sein können. Eine im außertextlichen Raum gefundene Gliederung konnte dann als Ordinatensystem für die Klassifikation der Texte selbst angenommen werden.

Die Kontrolle am Einzelmaterial müßte dann auf dem Fuß folgen, weil sonst die Gefahr der unfruchtbaren Spekulation allzugroß wäre. Nur in einem gegegenseitigen - zugegebenermaßen hermeneutischen - Absichern deduktiven und induktiven Vorgehens kann ein vorläufiger Erfolg unseres Unternehmens wenigstens als möglich erscheinen, denn nur auf diese Weise kann man mit einiger Sicherheit grundsätzliche und irreparable Unvollständigkeiten eines Systems vermeiden, ohne sich allzusehr der Gefahr eines Schematismus auszusetzen.
Auch die folgenden Ausführungen werden sich an diesen theoretischen Grundsatz halten, d.h. zunächst versuchen, außertextliche Konstituenten zu finden, und dann sich darum bemühen, denkbare oder tatsächlich vorhandene Texttypen den gewonnenen Kategorien beispielsweise und unverbindlich zuzuordnen.
Nun gibt es ja tatsächlich eine ganze Reihe von Versuchen, die Fülle vorhandenen Schrifttums zu gliedern und zu klassifizieren. Am bekanntesten sind die poetologischen Bemühungen seit Aristoteles, am umfassendsten die aus praktischen Gesichtspunkten vorgenommenen Einteilungen in Bibliotheken, Archiven und Bücherkatalogen. Neben soziologischen finden sich in letzter Zeit semiotische und linguistische Untersuchungen, welch letztere sich bemühen, die außersprachlichen Komponenten so vollständig wie möglich zu ignorieren, die Texte selbst nur noch unterschiedlichen sprachlichen Konstituenten abzufragen.

Zu den herkömmlichen poetologischen Arbeiten ist in unserem Zusammenhang zu bemerken, daß sie neben ihrer Unvollständigkeit vor allem ihrer zu globalen Differenzierung wegen kein geeignetes, weder theoretisches noch praktisches Grundschema anbieten. Jüngere Arbeiten wie die Käte Hamburgers [1] zeigen das deutlich: sie vermögen nicht auch nur annäherungsweise eindeutige Abgrenzungen zu geben zwischen lyrischen, epischen und dramatischen Texten, an welcher klassischen Einteilung sie weitgehend festhalten, auch nicht zwischen fiktiven und existentiellen. Dem Versuch einer Musterpoetik, wie sie Emil Staiger [2] andeutungsweise vorgetragen hat, müßte andererseits entgegengehalten werden, daß mit den von rein äußerlichen Merkmalen ausgehenden Kriterien für die herkömmlichen Gattungen nichts ausgerichtet ist und für un-erhörte

Formen kein Raum gegeben ist, weil eben nur diese herkömmlichen Gattungen als unbezweifelte Axiome gelten. Einen Schritt weiter dagegen könnte uns das bekannte poetologische Rad Julius Petersens[3] führen, das dieser im Anschluß an Bemerkungen Goethes erstellt hat: es gelingt ihm tatsächlich, durch Gegenüberstellung äußerlich-formaler, inhaltlicher und stilistischer Kriterien eine differenzierte Klassifizierung zu geben.



Der große Vorteil dieses Systems besteht gerade darin, daß Gesichtspunkte ganz verschiedener Observanz berücksichtigt sind. Da sie sich gegenseitig ergänzen und relativieren, kann Petersen mit allen denkbaren Kombinationen und unter Beachtung jeweils vorherrschender Komponenten zu einem alle bekannten Dichtungsformen seiner Zeit umfassenden Schema gelangen. Unter Einbeziehung einer Opposition fiktiv/nicht-fiktiv würden auch nichtpoetische Texte Eingang in sein System finden. Die Schwierigkeit besteht jedoch darin, daß die Anhaltspunkte, nach denen Zustand und Handlung, Darstellung und Bericht überhaupt und nach ihrer quantitativen Bedeutung in den einzelnen Texten bestimmt werden können, nicht festzustellen sind, ohne bereits Ergebnisse zu präjudizieren.

Dieser Problematik unterliegen in gesteigertem Maße die linguistischen Klassifikationsversuche neueren Datums. Roland Harweg[4] hat gezeigt, daß es einen Weg gibt, auch größere Spracheinheiten als den Satz mit Mitteln exakter Wissenschaft in den Griff zu bekommen, indem man mit Hilfe der von ihm ausgearbeiteten Substitutionsmethode Sätze und Satzgefüge zu einem Ganzen integrieren und von anderen Ganzen deutlich und linguistisch eindeutig unterscheiden kann. Seine Methode befreit uns zunächst von der bisher immer latent vorhandenen Aporie, daß Texte als solche nur mit Hilfe außersprachlicher Kriterien zu erkennen und zu bestimmen waren. Diese Methode, so theoretisch einwandfrei sie auch sein mag, erlaubt uns jedoch vorläufig nur eine sehr grobe Differenzierung, nämlich sachbezogene und nicht sachbezogene Literatur grundsätzlich auseinanderzuhalten und diese Unterscheidung am aktuellen, als emisch[5] erkannten Text laufend zu kontrollieren und dabei frequenzmäßige Unterscheidungen, damit eine Art Klassifikation aufzustellen[6].

Den Vorwurf, den gesamten Textkosmos in ein System einzuspannen, das zu grob und zu global, deshalb zu vage und unergiebig ist, kann man den an der Praxis ausgerichteten Gliederungen, wie sie in Bibliotheken und Bücherkatalogen verwendet werden, nicht machen. Bei ihnen, die ausschließlich vom Textinhalt ausgehen, kann die Untergliederung bis in aufs subtilste ausgearbeitete Verästelungen vorgenommen werden. Was dagegen unsre Absichten betrifft, so muß die grundsätzliche Offenheit des Systems als problematisch angesehen werden : Es gliedert nur das tatsächlich greifbare, nicht das potentielle Schrifttum, so daß niemals die Möglichkeit einer grundlegenden Lücke ausgeschaltet ist. Da wir es - praktisch gesehen - natürlich auch nur mit de facto vorhandener Literatur zu tun haben, muß dieser Einwand an die zweite Stelle rücken. Wichtiger ist, daß nach inhaltlichen Gesichtspunkten vorgenommene Klassifikationen zwar die Gewähr einer relativ ausführlichen Unterteilung vor allem der Sachtexte bieten, dagegen nicht gewährleisten, daß die gewonnenen Unterscheidungen eine Entsprechung in der sprachlichen Differenzierung haben, da der Inhalt einen sicherlich nur zweitrangigen Einfluß

auf die sprachliche Gestaltung im einzelnen hat, wie vergleichender Blick auf politische, historische, juristische Abhandlungen etwa auf Anhieb zeigt[7].

All die bisher vorgeführten Klassifikationsbemühungen kranken, wie ich glaube, vor allem daran, daß sie ihre Kriterien ausschließlich aus den Texten selbst ableiten, damit entweder Ergebnisse vorwegnehmen, u.U. sogar verfälschen, oder dem potentiellen Gesamtbestand in ungenügender Weise gerecht werden. Mir scheint vom theoretischen Standpunkt aus nur ein anderer Weg begehbar zu sein, nämlich eine texttranszendente Position einzunehmen, einen archimedischen Punkt zu beziehen, von dem aus ein etwaiges Gesamtsystem zu überschauen ist. Wenn uns das gelingt, gilt es freilich zu überprüfen, ob die in den genannten traditionellen Gliederungen verwendeten, zweifelsfrei wesentlichen Gesichtspunkte ihren angemessenen Niederschlag in unserer Einteilung gefunden haben.

Nun bietet sich ein solcher archimedischer Ort an, wenn man bedenkt, daß zur Kommunikation neben dem Text selbst Sender und Empfänger gehören. Ist der Text als Ansatzpunkt unzureichend, könnte man es mit dem Empfänger versuchen. Es gibt sicherlich ein Primärpublikum, für das ein Text zunächst bestimmt ist und das hinsichtlich Quantität und Qualität, nach gruppen- und individualpsychologischen, nach soziologischen und anderen Gesichtspunkten differenziert werden könnte[8]. Jedoch erscheint es als unmöglich, diese primären Leserschichten aus der amorphen Masse der potentiellen Leserschaft, also der ganzen analphabetischen Sprachgemeinschaft, herauszukristallisieren, auf daß eine gültige Klassifikation entstünde. Man wird aber auch diesen Aspekt im Auge behalten müssen, immerhin spielt das Publikum eine nicht zu unterschätzende Rolle bei der sprachlichen Gestaltung eines Textes, auch schon in der Intention des Text-Verfassers, dem wir uns nun zuwenden wollen.

Der sprachwissenschaftliche Allgemeinplatz, daß langue sich nur in der parole realisiert, parole aber nur durch das Sprachsubjekt, berechtigt dazu, den Textschreiber unter Außerachtlassung der bisher erwähnten Gesichtspunkte zum

Fundament einer relativ vollständigen und praktikablen Textklassifikation zu mach
Dabei wären für eine Untersuchung des jeweiligen Individualstils alle Gesichts-
punkte aufzuführen, die Steger bei seiner Typologie der gesprochenen Sprache au
gestellt hat. Die überaus vielfältigen und ausführlichen Differenzierungen, die da
zur Folge hätte, sind aber für uns nicht durchführbar und auch gar nicht notwend
behaupten wir doch, daß in der geschriebenen Sprache aufgrund ihres größeren
Wirkungsgradius im Räumlichen und Zeitlichen, ihrer jederzeitigen Kontrollierbar-
keit, ihrer Bindung an ein überkommenes Schriftsystem, insgesamt also ihrer stär-
keren Bezogenheit auf subjekttranszendente Gegebenheiten die Bedeutung der per-
sönlichkeitsbedingten Merkmale zurücktritt hinter die nach außen orientierte Ab-
sicht des Schreibers, also seinen mehr oder weniger bewußten Intentionen bei der
Abfassung. Es muß nicht betont werden, daß der Individualstil dabei seine grund-
sätzliche Bedeutung nicht verliert, für eine allgemeine Texttypik aber außer Ach
bleiben kann und muß.

Wir gehen also von der Voraussetzung aus, daß, wie immer ein Text auch deter-
miniert sein mag durch inhalts-, gattungs- oder publikumsbedingte Regulative,
das formulierende Subjekt mit seiner Intention den Anstoß und die Richtung gibt,
in der eine wie immer geartete Texteinheit sprachlich gestaltet wird. Das Subjek
gibt uns die Kriterien in die Hand, Texteinheit überhaupt als solche zu erkenner
sei es durch äußere Merkmale, wie Überschrift und Schlußzeichen, durch erkenn
baren Anfang nach einer Leerstelle auf dem Papier o.a., sei es durch imma-
nente Merkmale, wie sie Roland Harwegs subtile Substitutionsmethode nachgewie:
hat.

2. Versuch einer Typik der geschriebenen Sprache entsprechend der Schreibintention

Eine erste und grundlegende Gliederung bietet sich an, wenn wir, wieder im
Hinblick auf das bekannte Kommunikationsdreieck, die Subjektintention interpre-
tieren. Sie kann orientiert sein am Objekt, am Kommunikationspartner oder nur
an sich selbst. Es läßt sich also leicht eine Oppositionsreihe von Relationen auf
stellen : ich-es (A)/ich-du (b)/ich-ich (C). Diesen Relationen, als denkbare

Grunddispositionen der Schöpferintention verstanden, würde auf der Seite der daraus hervorgehenden verschiedenartigen Textmöglichkeiten ungefähr die Reihe Sachtext/Mitteilung/Reflexion (im weitesten Sinn) entsprechen[9]. Isoliert genommen bieten diese Typen natürlich noch keine befriedigende, ausreichende und praktikable Lösung, und es soll hier ein für allemal betont werden, daß eine solche Gliederung keinen distinguierenden Charakter beansprucht, sondern gesetzt ist als Schema der für die Textkonstitution überwiegend verantwortlichen Motivationen. Jeder Text trägt a priori, sofern er das Licht der Öffentlichkeit erblickt, Mitteilungscharakter (sei dieser objektbedingt oder expressiv), jeder Text ist sachgebunden (Sprache ist nun einmal informationshaltig, sehen wir zunächst einmal von dadaistischen Zerstückelungen ab), und jeder Text ist in irgendeiner Weise Ich-Aussage (selbst ein Telefonbuch gibt durch Anordnung, Fettdruck, Informationsredundanzen und dgl. einen gewissen Einblick in das Aussagesubjekt, nämlich das der Gruppe Telefonbenützer). Es steht uns hier also möglicherweise eine Kriterienreihe zur Verfügung, deren Glieder als Konstante und Variable eines Gesamtschemas zur Anordnung des Textkosmos dienen können. Hat sie doch den in der Poetik verwendeten Prinzipien gegenüber voraus, leichter und eindeutiger abgrenzbar zu sein. Ein konkreter Text wäre hinsichtlich seiner klassifikatorischen Zuordnung abzufragen, 1. ob eine Leserresonanz primär vorgesehen ist, 2. ob, wenn das nicht der Fall ist, der behandelte Gegenstand in unmittelbarer und unabdingbarer Relation zu einer empirisch und außerhalb des Textes (bzw. des Schreibers) vorhandenen Wirklichkeit steht oder umgekehrt nur auf die (innere) Wirklichkeit des Textes bzw. des Schöpfers verweist. Ein grundsätzliches Problem, das sich der Poetik immer wieder gestellt hat, haben wir auf diesem Wege impliziert, nämlich fiktive und nicht fiktive Texte unter einem übergeordneten Gesichtspunkt unterscheidend einzuordnen. (Unter fiktiv sei dabei jede Art der Aussage verstanden, die nicht primär an eine Kontrolle durch die empirische Außenwirklichkeit gebunden ist.)

Die angedeutete Trichotomie bedarf nun allerdings einer weitgehenden Differenzierung. Nachdem unsere Konzeption ja von der Subjektsintention ausgeht, hat entsprechend die oben aufgestellte Oppositionsreihe eine Konstante 'Ich' , die mit drei Alternativvariablen gekoppelt werden kann, nämlich 'es' , 'du' und 'ich'. Eine zu erweiternde Gliederung müßte also versuchen, den Bereich der Konstanten aufzufächern. Es stellt sich füglich folgende Frage : Wie kann die Ich-Intention gegenüber ihren drei Kommunikationskomplementen differenziert werden?

Wir müssen versuchen, möglichst alle denkbaren subjektiven Verhaltensweisen der Wirklichkeit gegenüber zu erkennen, du-, ich- und es-bezogene. Damit berühren wir Probleme der Semiotik : Wie reagiert der Mensch auf die Wirklichkeit, zu der in unsrem Sinn natürlich nicht nur die empirische Faktenwelt, sondern ebenso die subjektive Erlebniswelt, theoretische Abstraktionen, allgemeine Gesetzmäßigkeiten u.s.w. gehören.

Die entsprechenden Forschungsergebnisse lehren uns, daß es primär 5 Reaktionsmöglichkeiten gibt, nämlich die der Identifikation (0), der Charakteresierung (1), der Wertung (2), der Aktionsauslösung (3) und der relativierenden, bzw. differenzierenden Zusammenschau (4). Praktisch und etwas vereinfachend könnte man sagen, der Mensch reagiert auf die Wirklichkeit mit den Fragen wann/wo; was; wie; wozu und in welchem Zusammenhang.

W.C. Morris, dessen Buch 'Signs, Language und Behavior' [10] ich diese Zusammenstellung entnommen habe, zeigt, wie diesen Verhaltensweisen bestimmte sprachliche Zeichen zugeordnet sind, sog. Assignators (im einzelnen nennt er sie identifiors, appraisors, designators, prescriptors und formators), so daß wir auch damit rechnen dürfen, daß eine Texttypologie, die nach diesen Verhaltensweisen ausgerichtet ist, auch sprachliche Differenzierungen impliziert. Und wenn diese Verhaltensweisen tatsächlich die einzig beobachtbaren, vielleicht sogar denkbaren sind, dann

müßte in ihnen ein vollständiges System auch von Texten, soweit sie von der Intention her zu gruppieren sind, sich gespiegelt wiederfinden.

Ein zunächst noch rein theoretischer Versuch der Koppelung der von mir vorgeschlagenen Einteilungsprinzipien würde folgendes Schema ergeben:

A sachorientierte	B du-orientierte(r)	C ich-orientierte Erlebnis
1. Analyse	Information	Bestandsaufnahme
2. Beurteilung	Überredung	Bewertung
3. Anwendbarkeits-prüfung	Appell	Auswertung
4. Systematisierung	Belehrung	Reflexion
5. (ungezieltes Ex-periment)	Unterhaltung	'Gedanken'spielerei

Über die Problematik dieses Schemas hinsichtlich seiner Spartenausfüllung im einzelnen bin ich mir vollauf bewußt. Es kommt aber, wie gleich deutlich werden wird, auf diese einzelnen Benennungen gar nicht an. Eine kurze vervollständigende Interpretation soll das zeigen. Vorab sei nur erwähnt, daß die Einbeziehung der Identifikation (0) in unser System nicht notwendig ist, weil sie nur als die Kenntnisnahme eines Gegenstandes gemäß seinem bloßen Vorhandensein in einer Ja-Nein-Entscheidung zu verstehen ist und wir uns nur für den Ja-Entscheid interessieren können. Zu betonen ist zweitens, daß alle vorgeschlagenen Rubriken zunächst völlig unabhängig von irgendwelchen Inhalten nur als Differenzierungen von Intentionen gemeint und zu verstehen sind. Daraus ist es zu erklären, daß sich die Ausfüllungen der

Kolumnen A und C prinzipiell so ähnlich sind. Sie müssen sich gleichen, weil [illegible] dieser von Konkretisierungen absehenden Benennungen das ich (bzw. der Text selbst) in C natürlich objektiviert ist, zur Sache werden, dem gegenüber das Subjekt-Ich reagiert. Weil aber im Hinblick auf die anschließend zu [illegible] konkreten Textreinheiten der Unterschied von ich- und es-bezogener Aussage außerordentlich wichtig ist, habe ich mich von vornherein für unterschiedliche Bezeichnungen entschieden. Drittens fällt auf und ist als Konstitutivum zu berücksichtigen, daß die künstlich auseinandergehaltenen Einzelaspekte selbstverständlich nie absolut isoliert vorkommen und zudem keineswegs gleichwertig zueinander stehen, vielmehr in einer hierarchischen Stufenfolge angelegt sind von 1 - 4 aufsteigend: 1 wird immer in 2, 1 und 2 in 3, 1,2,3 in 4 vertreten sein müssen; da aber an jeder Stelle der Hierarchie haltgemacht werden kann, müssen die Stufen einzeln vermerkt werden, und die Einteilung sagt, wie schon erwähnt, nur etwas über die dominierende Intention aus. Viertens muß nochmals darauf hingewiesen werden, daß die Grenzen zwischen A, B und C durchaus offen sind. B steht füglich in der Mitte, weil sowohl es- wie ich-bezogene Intentionen nach Ausdruck streben und dies eben gegenüber einem Du geschieht. Eine Fülle von Differenzierungsmöglichkeiten eröffnet sich also. Eingangs wurde bereits darauf aufmerksam gemacht, daß natürlich auch A und C in gegenseitiger Verfugung zu sehen sind. Ein schwieriges Problem stellt die augenfällige Unvollständigkeit des Schemas dar. Kann doch ein wesentlicher Teil des denkbaren Schrifttums vorläufig nicht angemessen berücksichtigt werden, solange der wesentliche Apsekt der s c h e i n b a r e n Intentionslosigkeit nicht seine entsprechende Stelle eingenommen hat, wobei unter Intentionslosigkeit die Möglichkeit des zweckfreien Spiels gemeint ist. Da es sich dabei aber dennoch eindeutig um eine Intention handelt, eben die der Zweckfreiheit, scheint es mir - trotz der theoretischen Unvereinbarkeit mit den genannten vier Verhaltensweisen - erlaubt zu sein, sie in einer fünften Sparte anzufügen. Dabei muß man sich bewußt bleiben, daß die oben angedeutete Hierarchie nicht fortgesetzt wird, sondern daß man mit grundsätzlich andersartigen Faktoren rechnen muß.

Weitere Modifikatoren ließen sich, wie dargetan, in ziemlich ungemessener Zahl

rwägen, vor allem jene wichtigsten, ob eine Differenzierung nach Inhalt, Gattung und Adressat sich als notwendig erweist, was uns am Ende dieses eferates noch beschäftigen wird. Zunächst würde unser Schema sich bei olchen Rücksichten ins Unüberschaubare ausweiten.

a, unser System selbst in seiner jetzigen Form läßt schon eine immense Viel- alt zu, wenn man bedenkt, daß die 14 Typen schon geordnet werden müßten ntsprechend der Rangordnung in der Reihenfolge ihrer Bedeutung pro Text, lso A 1234 gegen A 1342 gegen A 1432 u.s.w.

)iese mögliche theoretische Vielfalt ist keineswegs ein Einwand gegen unseren nsatz, sie bestätigt ihn insofern, als sich ja hier die tatsächliche Unüberseh- arkeit möglichen Schrifttums spiegelt. Für unsern praktischen Zweck dagegen uß gefragt werden, wo eine derart unbegrenzte Unterteilung aufhört, prakti- abel und auch notwendig zu sein. Es kommt ja nur darauf an, Haupttypen fest- uhalten, die denkbare Subspezies im wesentlichen vollgültig repräsentieren.

. Anwendung

n dieser Stelle sei nochmals darauf hingewiesen, daß wir bisher ausschließlich heoretisch denkbare Intentionen für die Entstehung von Texten betrachteten. Eine inwandfreie Zuordnung konkreter Texte müßte sich der faktoriellen Analyse be- ienen, um je nach Anzahl einschlägiger Faktoren pro Text Gruppen, dann ielleicht auch Typen und Klassen zu ermitteln.

)och kommt es bei unserem Unternehmen nicht darauf an, beliebige konkrete exte irgendwie zu typisieren. Vielmehr wollen wir ja umgekehrt durch die ypik erreichen, den gesamten Textkosmos repräsentativ zu erfassen auf eine Veise, die verspricht, daß den einzelnen konkreten Texten nicht allzusehr Ge- alt angetan wird. D.h., das System muß deduktiv vorweggenommen sein und eine Zureichendheit und Effektivität dadurch erweisen, daß 1. keine Lücke vor- anden ist (jede der herkömmlich bekannten Textsorten muß ebenso wie jede un- ekannte, aber denkbare ihren Platz haben) und 2. die Sparten sinnvoll ausge- üllt sind (die theoretisch-abstrakten Unterschiede müssen ihre Entsprechung in der

empirischen Wirklichkeit der vorhandenen Texte haben). Sofern das System, wie oben angedeutet, tatsächlich a l l e möglichen Intentionen umfaßt, dürfte der ersten Bedingung Genüge getan sein. Zur Prüfung der zweiten Voraussetzung sollen die folgenden Ausführungen dienen. Sie versuchen zu zeigen, wie unser Schema gefüllt werden kann mit Textsorten, die wir als voneinander unterschieden schon kennen.

So würde sich etwa folgende Aufstellung ergeben:

A 1 Reporte, Regesten, Kataloge, die nicht im Hinblick auf ein Publikum geschrieben sind.

A 2 Resumees, Memoranden, Polizeiberichte

A 3 Gebrauchsanleitungen, Funktionsbeschreibungen

A 4 Abhandlungen

A 5 entfällt, da zweckfreie, sachorientierte Spielerei nicht schriftlich fixiert wird.

Bei einer Aufzählung der B-Typen stellt sich am stärksten die Schwierigkeit in den Weg, daß es ich- und es-lose Mitteilungen nicht geben kann. Die folgenden Beispiele können also nur annähernd stehen, sie sind ausgesucht unter dem Gesichtswinkel, daß die Mitteilungskomponente die beiden andern so weit überwiegt, daß der in Frage stehende Text als ohne das Du nicht entstanden gedacht sein kann.

B 1 die unkommentierte Benachrichtigung, also z.B. Zeitungskurznachrichten, evtl. auch Telegramme, die ihre Kürze freilich nicht so sehr der isoliert du-orientierten Intention, sondern der Rücksicht auf den Geldbeutel des Schreibsubjektes verdanken.

B 2 die einfachen Formen der Werbung und Reklame, also v.a. der Anpreisungsslogans, dann auch Inserate und Rechnungen, die zunächst einmal ihren Gegenstand charakterisieren, wobei eine direkte Aufforderung zum Kauf oder Verkauf sich nur sekundär ablesen läßt. Man könnte sie freilich auch zur nächsten Gruppe rechnen.

3 3 Aufrufe, Vorschriften, Anweisungen, Verbote

3 4 alle Arten von Lehrschriften

3 5 Rätsel, Witz

Bei einer Betrachtung von C nun wieder ist darauf zu achten, die Mitteilungskomponente weitestgehend auszuschalten. Konsequenterweise dürften wir also nur private, nie zur Veröffentlichung auch im engsten Kreis gedachte Texte heranziehen. Eine solche provisorische Konsequenz erlaubt uns allerdings die Aufnahme bisher vielfach vernachlässigter oder überhaupt nicht beachteter Typen, nämlich alle Vorstufen einer künstlerischen oder philosophischen Produktion.

Es würden sich also ergeben

C 1 alle Arten bloßer Aufzeichnungen von Erlebnissen, Gedanken u.s.w., also v.a. Notizbücher, Tagebücher

C 2 Zettelkästen, Vornotizen, Stoffsammlungen

C 3 Konzepte und Gliederungen

C 4 fertige Skizzen und Vorstudien

C 5 Nonsensaufzeichnungen

Es ist augenfällig, daß mit den aufgeführten Typen die Fülle des Materials auch nicht annäherungsweise bewältigt ist. Da aus erwähnten Gründen eine faktorielle Zuordnung weder möglich noch auch sinnvoll zu sein scheint, versuchen wir diesem Dilemma wenigstens insoweit Rechnung zu tragen, daß wir neben den sozusagen reinen Formen den denkbaren Kombinationen der Kolumnen unser Augenmerk widmen und dabei die jeweiligen Prävalenzen der gekoppelten Komponenten beachten. Auf diese Weise kann sich unser System in nahezu beliebigem Umfang, der sich nach der praktischen Notwendigkeit richten muß, erweitern und differenzieren lassen.

Eine besondere Schwierigkeit unserer Aufstellung besteht darin, daß entsprechend der oben angedeuteten Verhaltenshierarchie die weitaus überwiegende Anzahl möglicher Texte, sofern sie nicht 'zweckfrei' konzipiert sind, der Sparte 4 angehören, weil kein Gegenstand so völlig isoliert und keine Intention so völlig rein gestaltet werden kann, weil zudem auch das Du, das Publikum also, einen größeren Zusammenhang

geboten bekommen muß oder diesen aus einem implizierten Vorverständnis selbst herzustellen hat, ein Vorwissen, das irgendwie in der Tiefenstruktur des Textes mitgegliedert oder abgerufen werden muß. Ja selbst die primitivste Aufzählung von Erlebnissen und Gedanken geschieht ja bereits innerhalb eines Kosmos, nämlich des integralen Ich. Für die folgende Aufzählung muß also berücksichtigt werden, daß streng logisch gesehen B und 4, meist also B 4 vorausgesetzt werden müssen, daß sie aber in vielen Fällen in so schwachem Maße konstituierend sind, daß sie für eine primäre Zuordnung vernachlässigt werden dürfen. Eine praktisch ausfüllbare Gliederung kann nur nach dem Prinzip einer Beachtung möglicher Prävalenzen einzelner Intentionen untereinander funktionieren. Bei einer Koppelung von A und B würden sich etwa folgende Konkretisierungen ergeben :

A1-B1 Versuchsanordnungen, Theaterprogramme, Telefonbücher
A2-B1 Zeugenberichte, Bulletins
A3-B1 Gebrauchsanweisungen, Rezepte
A4-B1 Abhandlungen, Referate

A1-B2 Gesetzestexte, evtl. Verträge
A2-B2 Verkaufskataloge
A3-B2 Bewerbungsschreiben, evtl. Propagandaschriften
A4-B2 Leitartikel, Kommentare

A1-B3 (milit.) Lageberichte, die meisten Geschäftsbriefe
A2-B3 Predigten bzw. entsprechende Traktate
A3-B3 Dienstanweisungen, Tagesbefehle u.s.w.
A4-B3 Manifeste, Parteiprogramme u.dgl.

A1-B4 Kurzfassungen, Abrisse (z.B. Lausberg, Ploetz)
A2-B4 Lehrbücher
A3-B4 Repetitorien
A4-B4 Kompendien

1-B5 Features

2-B5 Glossen, Causerien

3-B5 für diese Sparte kenne ich keinen Sammelbegriff, meine aber derartige Erzeugnisse wie die bekannten 'launigen Kochbücher' u.dgl., also unterhaltende Anweisungstexte

4-B5 Essays, unterhaltende Sachbücher

ür die folgende Aufstellung gilt die oben gemachte Einschränkung vor allem. ei allen poetischen Texten, also den ich-ich/text-ausgerichteten, gilt, daß wir ur das Endprodukt (seine Vorstufen hatten wir oben schon beachtet) modifizieren der klassifizieren können, weil die Vorstufen eben nicht du-orientiert waren. Nur von C4 und C5 aus lassen sich Kombinationen mit B vornehmen, was folgende Unterscheidung ergibt:

C4-B1 Autobiographische Bemerkungen, die meisten Privatbriefe

C4-B2 Aphorismen, Parabeln, Gebete

C4-B3 Fabeln, Gleichnisse u.s.w.

C4-B4 dokumentarische Dichtung, religiöse, ethische, metaphysische Abhandlungen, Autobiographien

C4-B5 nicht sachgebundene, sog. existentielle Dichtung

In den Bereich von C5 in seinen verschiedenen Koppelungen mit B fällt die ganze Sphäre des Grotesken mit allen seinen Spielarten, von den Dodererschen Kürzestgeschichten (C5-B1, denkbar auch C5-B5) über z.B. die Keunergeschichten (C5-B3) zu Kafka und dem Surrealismus (C5-B4). Die Verbindung C5B5 bezeichnet den ganzen Bereich der Nonsensliteratur bis hin zu spätdadaistischen Gebilden eines Mon oder Gomringer.

Nur kurz sei darauf eingegangen, welche Differenzierungen sich ergeben würden, wenn die nach der vorwiegenden Bedeutung der Komponenten angesetzte Reihenfolge umgekehrt wird, so daß jeweils die du-orientierte B-Komponente an erster Stelle steht:
An die Stelle von Zeugenberichten würde die Reportage treten (B1-A2), das Feuilleton würde den Essay ersetzen (B5-A4) u.s.w. Vorbereitende Untersuchungen müßten

ergeben, ob eine derartige Aufschlüsselung in jedem Fall sinnvoll ist, oder ob nicht vielmehr die Prädominanz der Du-Orientieru , also vornehmlich der Belehrung und Unterhaltung, sich in einer einheitlichen und also grundsätzlich festlegbaren Abweichungsrichtung sprachlich spiegelt. Das würde bedeuten, daß nicht unser ganzes System nach Alternativen abgetastet zu werden brauchte, sondern einige wenige Beispiele als Belege für mögliche Spielarten genügten.

Am einschneidendsten dürfte sich eine solche Gewichtsverlagerung der Komponenten im Bereich der Kombination CB bemerkbar machen, weil hier unter anderem der ganze Komplex der Unterhaltungsliteratur anfällt. B4-C4 wäre die Rubrik für Lehrdichtung und B5-C5 eben für die ganze Skala der Trivialliteratur von Hans-Ulrich Horster bis Hans Habe, vom Groschenheft bis zum anspruchsvollen Kriminalroman, von der Räuberpistole zum Boulevardstück. Die Umkehrung von C5-B1-4 ergäbe den Komplex der kabarettistischen Texte, während B5-C5 in der Clownerie repräsentiert wäre.

Unser Gesamtsystem hat sich vor allem im Bereich C außerordentlich reduziert. C5-B4 bzw. B4-C5 scheint ein Sammelbecken zu sein, das wegen seiner Überbelastung anscheinend eine Untergliederung erforderlich macht. Ich meine jedoch, daß hier eine weitere Differenzierung nicht unbedingt nötig zu sein braucht. Soweit ein unkontrollierter Überblick überhaupt eine Aussage erlaubt, läßt sich sagen, daß innerhalb der Gruppe C4 sprachliche Unterschiede auf das Konto der Individualstile, nicht aber klassifizierbarer Untertypen gehen. Ich bin mir bewußt, im Unkontrollierbar-Spekulativen zu bleiben, aber meine Beobachtungen haben ergeben, daß eine Mannsche Novelle in den sprachlichen Einzelerscheinungen bis hinauf zur Satzgruppe keine Unterschiede zu einem Roman von ihm aufweist, eine Böllsche Kurzgeschichte nicht anders aussieht als eine Erzählung des gleichen Autors, es sei denn sie habe dokumentarischen Charakter, in welchem Falle sie ja auch einer anderen Gruppe angehört.

Dies gilt, wie ich glaube, in gewisser Hinsicht selbst für die bekannten Großgattungen, Lyrik, Epik und Dramatik. Für das Drama wage ich zu behaupten, daß es sprachlich entweder nicht von einem epischen Text der gleichen Kategorie

abweicht (die Besonderheiten des Versdramas - etwa Peter Weiß - sind dann entsprechend der Lyrik vergleichbar), oder aber unter einem ganz anderen Gesichtspunkt, nämlich dem der gesprochenen Sprache, betrachtet werden muß. Da beides fast immer in sehr komplizierter gegenseitiger Verschränkung auftreten wird, die wir nicht ohne weiteres lösen können, sollten dennoch Hörspiele und Dramen unbedingt in entsprechender Anzahl zum Zwecke der Überprüfung und Materialgewinnung aufgenommen werden.

Was die lyrischen Formen anbelangt, so bin ich der Meinung, daß sie hinsichtlich ihrer sprachlichen Gestalt aufzufassen sind als 'Allotexte' zur Prosa, d.h. sie bilden zusammen mit dieser Texteme, innerhalb deren komplementäre Distribution herrscht. An allen, auch modernsten, lyrischen Gebilden läßt sich zeigen, daß sie entweder so frei und nur vom Individualstil geprägt sind, daß sie durch überhaupt keine Typik erfaßt werden können, oder daß sie sich überhaupt nicht von 'normaler' Prosa unterscheiden außer in der zeilenmäßigen Anordnung, oder aber Sprach- und Stil-alternativen zur Prosa bieten, die nach festen Regeln genormt sind, die immer gelten. Solche Regeln betreffen die Satzanordnung, die Wortstellung, den Tempusgebrauch und noch wenige weitere Erscheinungen. Sie sind also überschaubar und meist bedingt durch Versbau, Reim und Rhythmus. Auch hier müßte natürlich eine endgültige Prüfung am Material erfolgen.

Wie dem aber auch sei: Was immer eine Kontrolle ergeben sollte, es besteht ja immer die Möglichkeit, für etwaige nachweisbare gattungsbedingte Modifikationen in der C Gruppe Unterteilungskriterien aus der herkömmlichen Poetik zu verwenden, falls nötig, dann am sichersten aus der subtilen Klassifikation von Petersen.

Wie Sie bemerkt haben werden, sind wir jetzt unversehens wieder aus dem Bereich des Schreibe-Subjekts und seiner Intentionen herausgeraten. - Anfangs hatten wir ja offen lassen müssen, wie weit gattungs-, inhalts- und leserbedingte Terminanten einen Text bestimmen.

An dieser Stelle erweist sich m.E. nochmals die Anwendbarkeit unsres Systems. Es stellt sich nämlich heraus, daß diese drei Modifikationsmöglichkeiten nahezu ausschließlich unseren drei Hauptgruppen A B C zuzuordnen sind. Eine inhaltliche

Abhängigkeit tritt bei den vornehmlich sachorientierten Texten als latente Terminante in Erscheinung, die gattungsbedingte bei den ich/ text-orientierten und die leserbedingte bei den Du-orientierten. Es müssen also, wenn überhaupt, nicht immer alle drei als Variable auftreten. Die hervorstechende Komponente innerhalb unserer kombinierten Typen bestimmt eine allenfalls notwendige subjekttranszendente Modifikation.
Falls sich also von den Aussage-Inhalten her ergeben sollte, daß etwa ein naturwissenschaftlicher Essay anderen Bedingungen im Sprachlichen unterliegt, dann müßte gefragt werden, ob die praktisch gehandhabte Großgliederung in naturwissenschaftlich exakte und geisteswissenschaftlich hermeneutische Typen sich als ausreichend erweist, was anzunehmen ist. Ansonsten liefern die bekannten kulturhistorischen, praktischen oder philosophischen Schemata ausreichende Unterteilungen.

Am wichtigsten scheint mir aber die Frage nach dem Leser zu sein. Gemäß dem Stegerschen Abriss ließen sich eine große Zahl von Variablen dem B-Komplex aufsetzen. Wie weit man in der Subtilität der Unterscheidung gehen muß, kann wieder nur die Praxis entscheiden. Es ist aber ein Faktum, daß sich sprachlich niederschlägt, ob ein Kinderlehrbuch oder ein Lehrbuch zur Quantenmechanik ar gefertigt wird, ob die Bildleserschaft oder die von 'Poetica' oder den 'Sinologischen Mitteilungen' angesprochen wird, ob man für den 'Eichstätter Bistumsbote oder den 'Feuerreiter' schreibt. Auch hier ist ja die Möglichkeit denkbar, daß solche etwaigen Modifikationen auf einen einfachen, ein für allemal gültigen Nenner gebracht werden könnten, daß also wiederum nur wenige Beispiele repräsentativ sein können. Da das allem Erwarten nach aber nicht der Fall sein w muß man sich m.E. praktisch so behelfen, daß bei jedem für das Corpus auszuwählenden Textexemplar die Frage gestellt wird, ob und in welchem besetzungsmäßig relevanten Maße Alternativen denkbar und vorhanden sind.

Ich habe bei meinen Ausführungen offen lassen müssen, wie groß nun endgültig repräsentatives Corpus sein müsse, einfach weil in vielen Fällen eine Prüfung an Material notwendig ist. Dennoch läßt sich sagen, daß der Typenapparat, sofern

überhaupt mein Ansatz richtig war, nicht unter 15 Rubriken haben darf und für praktische Zwecke nicht mehr als höchstens 100 Rubriken zu haben braucht.

Ich bin mir, um das abschließend nochmals zu betonen, der Fragwürdigkeit und Diffusität meines Unternehmens durchaus bewußt und kann nur nochmals an die eingangs geschilderte allgemeine Problematik erinnern, deren Grad im Verlauf des Referates vielleicht noch deutlicher geworden ist. Wenn aber meine Gedanken einer ausführlichen Diskussion Anregungen gegeben hätten können, wäre, glaube ich, wenigstens der Sache wesentlich geholfen.

Anmerkungen

1) Hamburger, Käte : Die Logik der Dichtung, Stuttgart 1957.

2) Staiger, Emil : Andeutung einer Musterpoetik, in Festschrift Kunisch, Berlin 1961, S. 354 ff.

3) Petersen, Julius : Zur Lehre von den Dichtungsgattungen, Festschrift Sauer, Berlin 1927, S. 72 ff.

4) Harweg, Roland : Pronomina und Textkonstitution, München 1968.

5) Unter emisch versteht Harweg einen aus inneren (syntaktischen, semantischen) Gründen als geschlossen zu verstehenden Text im Unterschied zu einem durch äußere Merkmale konstituierten etischen Text.

6) Immerhin ließe eine solche Frequenzrelation, statistisch genau, wie sie herzustellen ist, eine logisch einwandfreie Textklassifikation zu. Doch ist wegen ihres eindimensionalen Ansatzes zu vermuten, daß dabei die Erhellung sprachlicher Merkmale verschiedener Texttypen nicht ausreichend oder überhaupt nicht vorgenommen werden kann, da intentionale (inhaltliche wie gattungsmäßig-formale) Gesichtspunkte nicht berücksichtigt werden.

7) Daß die inhaltliche Komponente nicht völlig außer acht gelassen werden darf, versteht sich aus dem obigen von selbst; wir müssen darauf noch zurückkommen.

8) Vgl. hierzu Steger, Hugo : Gesprochene Sprache, in: Satz und Wort im heutigen Deutsch, Düsseldorf 1967, S. 259 ff.

9) Ich halte es nicht für notwendig, die Mukarowskische Ausweitung unseres an Bühler ausgerichteten Systems um die ästhetische Funktion einer ich-Text-Relation zu übernehmen. (Vgl. Mukarowski, Jan : Die poetische Benennung und die ästhetische Funktion der Sprache, in: Kapitel aus der Poetik, Frankfurt 1967, S. 44 ff.). Bei den reinen Orientierungstypen, die wir gerade betrachten, ist die appellative und darstellende Funktion,

auch wenn sie expressiver Natur ist, dem B-Typus zuzuordnen, während der C-Typus außerhalb jeden Mitteilungs- und ichtranszendenten Objektbezuges steht. Die Entstehungsintention von C-Texten (also vor allem von poetischen Texten) ist nicht an ihnen (den Texten) selbst orientiert, sondern rekursiv am Aussagewillen des Ich. Die Formungsintention freilich unterliegt wieder anderen Maßgaben, unter anderem denen, die der entstehende Text selbst fordert. Dies gilt jedoch in unterschiedlichem Maße bei allen drei Typen, so daß es als untunlich erscheint, hier eine eigene Sparte anzusetzen.

10. Morris, Signs, Language, and Behavior, New York 1946. Wir übernehmen Morris, semiotisches System, ohne von seiner problematischen Einteilung sprachlicher Typen Gebrauch zu machen. (Vgl. Harwegs Kritik a.a.O. S. 329 f.).

Über Corpusgewinnung und Dokumentation im Mannheimer Institut für deutsche Sprache

von Manfred W. Hellmann

Übersicht:

0. Einleitung (26)

1. Zum Begriff "Dokumentation" (28)
1.1. Negative Eingrenzung (28)
1.2. Zu Ziel und Zweck einer Mannheimer Dokumentation (29)

2. Zur Frage der Texterfassung (30)
2.1. Sachliche und mengenmäßige Analyse der definierten Sprache (30)
2.2. Praktische und theoretisch-systematische Gliederung der definierten Sprache (30)
2.3. Möglichkeiten der Einschränkung (31)
2.4. Statistische Zwischenprüfung (32)
2.5. Zur Frage der exemplarischen Titelauswahl (33)

3. Zur Textaufnahme (35)
3.1. Informationskarten, Informationskategorien (35)

4. Zum Ausbau der Dokumentation (36)
4.1. Basisinformationen (36)
4.2. Ausbau im Mensch-Maschine-Wechselverfahren (36)
4.2.1. Kumulatives Wörterbuch als Beispiel (37)
4.3. Dokumentation als eigenständiges Arbeitsprinzip (38)

5. Zur Organisation (39)

Anhang: Zusammenfassung des Mannheimer Corpus (41)
Mengenmäßige Verteilung des Mannheimer Corpus (44)

Anmerkungen (50)

0. Einleitung

Die folgenden Überlegungen[1] gründen sich auf dreierlei : auf einige schon veröffentlichte Hinweise zur Mannheimer Dokumentation[2], auf die jahrelange Diskussion unter den Mitarbeitern des Instituts über das in Mannheim aufgenommene Corpus und das Verfahren seiner Aufnahme[3] sowie auf eigene Erfahrungen beim Aufbau eines Corpus westlicher und östlicher Zeitungstexte in der Außenstelle Bonn[4].

Peter von Polenz leitete seinen Vortrag auf der Mannheimer Frühjahrstagung 1965 folgendermaßen ein :

"Am Beginn jeder wissenschaftlichen Arbeit steht die Sammlung und Aufbereitung des Quellenmaterials, dessen Quantität und Qualität den Erfolg im voraus sichert oder aber in Frage stellt."... Der Sprachforscher habe "... eine wohlüberlegte Auswahl zu treffen, wenn er der Gefahr entgehen will, nur den zufälligen Funden aus bevorzugter Lektüre oder den einseitigen Perspektiven seines Sprachbewußtseins nachzugehen."[5]. Von Polenz fordert eine angemessene Berücksichtigung gerade auch der "Mittelschicht der sachlichen Informationssprache ...", des "... öffentlichen Sprachgebrauchs in Politik, Verwaltung, Wirtschaft und Technik"[6] und warnt : "Wir sollten uns nicht der Gefahr aussetzen, daß spätere Linguistengenerationen auch unsere Arbeit bemängeln müssen mit dem Vorwurf, wir hätten - trotz moderner sprachwissenschaftlicher Einsichten, Methoden und Mittel - die häufigsten und alltäglichsten Erscheinungen unserer Sprache nicht berücksichtigt, nur weil sie uns zu selbstverständlich, stilistisch problemlos oder etwa normwidrig erschienen."[7]. "Eine repräsentative Quellenauswahl für die Gegenwartssprache sollte also neben der (ständig variablen) Liste der z.Z. für bedeutend gehaltenen Schriftsteller auch eine der sprachsoziologischen Wirklichkeit angemessene Menge von Texten aus Tageszeitungen erfassen"[8]. An anderer

Stelle weist von Polenz auf die " 'Diskussionssprache' " hin, wie sie sich etwa in Bundestagsprotokollen findet[9].

Sicherlich sind einige der hier zitierten Anregungen beim Aufbau des Corpus berücksichtigt. So enthält es z.B. Tageszeitungen, wenn auch - mit Ausnahme der Bildzeitung - jeweils nur die erste Seite zweier Zeitungen aus einem Zeitraum von 1 bis 2 Monaten. Auch wurde die sprachsoziologisch sicher nicht unwichtige Trivialliteratur (in drei Exemplaren) sowie juristische und naturwissenschaftliche Sachliteratur berücksichtigt. Dennoch - und das ist der Grund, weshalb ich die Ausführungen von Polenz' hier ausführlich zitiere - sind sie in einigen Kernpunkten auch heute noch aktuell; sie berühren sich mit den Fragen, die in den institutsinternen Diskussionen um das Corpus immer wieder aufgetreten sind.

Diese Fragen betreffen, grob gesagt, einmal die Zusammensetzung des Corpus und zweitens die Art seiner Aufnahme auf Datenträger. Unter Verzicht auf Einzelheiten lassen sich diese Komplexe vielleicht folgendermaßen gliedern:
Zur Frage der Zusammensetzung:

1. Sind alle sprachwissenschaftlich wichtigen Gattungen oder Texttypen im Corpus enthalten? Welche müßten mindestens vertreten sein?
2. Sind alle wichtigen Sachgebiete vertreten und welche sollten mindestens vertreten sein?
3. In welchem Verhältnis stehen die im Corpus vertretenen Gattungen und Sachgebiete mengenmäßig zueinander? Wie läßt sich eine mengenmäßig adäquate Vertretung erreichen? Woran wäre sie zu messen?

Und zur Frage des Aufnahmeverfahrens:

1. Bietet die vollständige Aufnahme verhältnismäßig weniger Titel Vorteile gegenüber einer teilweisen Aufnahme vieler Titel?
2. Wenn nein: wie müßte eine angemessene Auswahl beschaffen sein?

Hinter diesem nur angedeuteten Fragenkatalog stehen weitere, grundsätzlichere Fragen, und ihm müßten zahlreiche sehr spezielle folgen. Sie auch nur zu formulieren, geschweige denn sie zu beantworten, ist einem einzelnen kaum möglich, sicherlich nicht in einer so kurzen Zeit wie der, die mir zur Verfügung stand.
Dennoch scheinen mir sowohl grundsätzliche Überlegungen zur Dokumentation als auch Überlegungen zur konkreten Verwirklichung durchaus angebracht.
Wenn ich sie hier riskiere, sind freilich kein systematisch geschlossenes und methodisch abgesichertes Grundlagenreferat, auch keine"konkreten Folgerungen" und Vorschläge für die Mannheimer Dokumentation, wie sie laut Einladung vorgesehen waren, zu erwarten. Weder habe ich meine Ansichten durch Literaturverweise oder durch eigene Nachforschungen genügend absichern können, noch lagen mir die Vorarbeiten von Kern rechtzeitig vor, auf die ich mich hätte stützen müssen. Ich bitte daher, das folgende als "Denkübung", als persönliche Meinungsäußerung zu betrachten, die zur Diskussion beitragen und sie anregen soll.

1. Zum Begriff "Dokumentation"

Ich bin versucht, die Frage nach klassischem Muster so zu formulieren :
"Was heißt und zu welchem Ende treibt man Dokumentation?"

1.1. Beginnen wir mit einer negativen Eingrenzung :

1. Dokumentation ist nicht Textaufnahme auf Datenträger, auch nicht Aufnahme repräsentativer Auswahlmengen aus definierten Textmengen, sofern eine solche Textaufnahme betrieben wird.
2. Dokumentation ist nicht Datenverarbeitung, auch nicht Textverarbeitung im Sinne von Anfertigung von Registern, Indices und Wortlisten.
3. Dokumentation ist nicht Zettelkastenersatz, etwa wenn man sie mit dem Ziel betreibt, für bestimmte begrenzte Arbeitsvorhaben bestimmtes Material schnell bereit zu stellen. Dazu ist sie zu langsam, zu schwerfällig und zu teuer; sie ist in dieser Hinsicht jedem modernen Zettelkastenverfahren unterlegen, wobei ich zum Zettelkastenverfahren auch etwa Lochkartensortierung mittels Sortiermaschinen zähle[10].

Dokumentation ist also nicht "ancilla grammaticorum" oder "lexicographorum". Freilich ist sie auch alles dies oder kann alles dies sein, aber damit ist sie weder grundsätzlich beschrieben noch in ihrer Leistung erschöpft.

Es scheint mir am besten, wenn ich meine Auffassung von Dokumentation durch Skizzierung möglicher Verfahrensweisen beim Aufbau einer Dokumentation darlege. Natürlich bin ich nicht der Meinung, daß es sich dabei um den einzigen Weg dorthin handele, obwohl bestimmte Gesichtspunkte wohl für jegliche Form Mannheimer Dokumentation Bedeutung haben dürften - aber wie gesagt, es scheint mir ein möglicher und sinnvoller Weg.

1.2. Im Institut soll ein Corpus von ausgewählten Texten zusammengestellt und maschinell verfügbar gemacht werden, das sprachwissenschaftliche Untersuchungsmöglichkeiten schaffen soll, und zwar:

1. Untersuchungsmöglichkeiten für Aussagen nicht nur über die aufgenommenen Texte selbst, sondern über die deutsche geschriebene Sprache der Gegenwart oder doch - falls dies undurchführbar sein sollte - für Aussagen über bestimmte definierte Bereiche dieser Sprache, etwa: der gedruckten Prosasprache der Gegenwart;
2. Untersuchungsmöglichkeiten nicht nur für bestimmte, z.Z. aktuelle Fragestellungen, sondern für eine möglichst große Zahl verschiedener, jetzt vielleicht noch unbekannter Fragestellungen;
3. Untersuchungsmöglichkeiten nicht nur für den Hausgebrauch der Mannheimer Mitarbeiter, sondern theoretisch für jeden, der über die deutsche Gegenwartssprache im definierten Sinne wissenschaftlich arbeiten will;
4. Untersuchungsmöglichkeiten nicht nur für jetzt oder die nächsten zwei bis drei Jahre, sondern für einen Zeitraum von vielleicht zehn bis zwanzig Jahren.

Jede Einschränkung dieser vier Grundsätze schränkt die Verwendbarkeit und Relevanz des Corpus auf Dauer ein; mehrere Einschränkungen zugleich können es zu einer Haustextsammlung rein interner Verwendbarkeit herabdrücken[11].

2. Zur Frage der Texterfassung

2.1. Ausgangspunkt aller Texterfassung muß eine sachliche und mengenmäßige Analyse der angezielten Gesamtmenge (der definierten Sprache) sein - zum Beispiel der gedruckten Prosasprache der Gegenwart. Diese sachliche Analyse der definierten Sprache ist wohl das eigentliche Thema der heutigen Tagung und wird auch in den anderen Referaten behandelt; ich möchte mich daher auf wenige Hinweise beschränken.

Wichtig scheint mir für die sachliche Analyse zu sein, daß sie sich unbedingt objektiver und quantifizierbarer Begriffe bedient. Eine Quantifizierung d.h. eine genaue Feststellung der Mengen und der Mengenverhältnisse, ist für jede Stufe der Texterfassung bis zur fertigen Modellmenge unerläßlich, - ein Probelm, das offenbar bisher vernachlässigt worden ist.

2.2. Kein Bearbeiter ist jedoch in der Lage, sich selbst einen auch nur annähernd genauen zahlenmäßigen Überblick über die ungeheure Flut des in Frage kommenden Materials zu verschaffen und die Ergebnisse in ein von ihm selbst entworfenes sachliches Gliederungssystem einzuordnen[12]. Es bleibt daher kein anderer Weg als der, sich eines Systems zu bedienen, für das schon weitgehende Zählarbeiten geleistet sind.

Praktikabel scheint mir daher eine sachliche Einteilung zu sein, wie sie im Buchhandel oder in Bibliotheken zur Erfassung und Gliederung des gesamten Schrifttums üblich ist. Der Börsenverein des deutschen Buchhandels, aber auch die Leitung der Frankfurter Buchmessen und andere Stellen[13] geben Statistiken heraus, die diese sachliche Gliederung samt Mengenangaben enthalten. Durch solche und ähnliche Hilfsmittel kann man jedenfalls einen ersten Überblick über die jährlich erschienenen Mengen und ihre Verhältnisse zueinander gewinnen[14].

Freilich entspricht diese sachliche Gliederung wahrscheinlich nicht sprachwissenschaftlichen Anforderungen. Die in der Gliederung der Verlags- und Buchhandelskataloge vorhandene Mischung von drucktechnisch-formalen, werk-formalen und inhaltlichen Kriterien muß sicherlich ersetzt werden durch eine

literaturwissenschaftlich oder sprachwissenschaftlich zureichende Gliederung, etwa wie die hier von Kern vorgeschlagene. Wieweit eine solche Gliederung mit der vorgefundenen korrespondiert, d.h. wieweit die Mengen- und Mengenverhältnisangaben zu übernehmen sind, kann jetzt nicht übersehen werden. Es ist durchaus möglich, daß auf eine vollständige Umformung in Richtung auf ein wissenschaftliches Gliederungssystem - anders gesagt: auf eine vollständige Füllung des wissenschaftlichen Gliederungssystems mit konkreten Mengenangaben - verzichtet werden muß. Sicher werden sich erhebliche Schwierigkeiten bei der mengenmäßigen Bestimmung der Gattungen und Untergattungen, nach welchem System auch immer, ergeben, jedoch muß diese Arbeit so weit wie irgend möglich geleistet werden.

Ebenso sicher ist aber auch, daß wir auf erstaunliche Quantitäten treffen werden. Als Beispiel seien die Tageszeitungen erwähnt [15)] : Es gibt rund 1400 verschiedene Tageszeitungen in der BRD. Jede erscheint rund 300 mal im Jahr. Rechnet man als Durchschnittsumfang nur etwa 15 Seiten [15a)], so ergeben sich 4500 Seiten pro Zeitung, also 6,3 Millionen Seiten insgesamt. Geht man von nur 2000 Wörtern pro Seite aus, ergibt sich eine Gesamtmenge von über 12 Milliarden Wörtern pro Jahr. Für die übrigen Periodika kann etwa noch einmal die gleiche Menge angenommen werden. Welche Menge sich in den letzten zwanzig Jahren ergibt, ist leicht, aber auch unnötig zu errechnen. Denn diese Mengen adäquat in ein Corpus zu überführen, ist ersichtlich unmöglich. Für Romane, Trivialliteratur, Sachbücher, Werbebroschüren, Kochbücher, Gebrauchsanweisungen und dergleichen werden die Mengen zwar nicht so groß, aber immer noch deprimierend groß sein.

2.3. Infolgedessen wird nun entweder eine Beschränkung auf bestimmte Bereiche der gedruckten Prosa - die ich durchaus für diskutabel hielte - erforderlich, oder aber wir müssen Hilfsverfahren, d.h. Hilfskriterien zur Auslese einführen, um die Ausgangsmenge zu vermindern. Wahrscheinlich wird man beide Wege gleichzeitig einschlagen müssen.

Wichtig ist dabei, daß es sich um Kriterien handelt, die anwendbar sind, ohne daß die Texte selbst bekannt sein müssen.
Ein solches Kriterium für Bücher und auch für Zeitungen könnte etwa die Auflagenhöhe oder die Verkaufsziffer sein. Ich hielte es für vertretbar, eine Gesamtmenge im engeren Sinne zu definieren, die nur solche vielverkauften Werke, Zeitungen usw. enthält. Ein Mittel zur Eingrenzung wären etwa die Bestsellerlisten, Verlagsangaben, Statistiken. Solche vom Beobachter unabhängigen Kriterien bleiben auch bei den weiteren Untergliederungen und Unterteilungen anwendbar. Jede weitere Untergliederung ermöglicht weitere Reduzierungen und feinere Mengenbestimmungen. Ob auf der zweiten oder dritten Stufe schon Untergliederungen nach literaturwissenschaftlichen Kriterien möglich sind, möchte ich bezweifeln, da sie wahrscheinlich keine Handhabe zur Quantifizierung der betreffenden Untergattung bieten.
Das Ergebnis dieses Vorgehens wäre - bildlich gesprochen - ein Kastensystem, in dem jeder Kasten mit mehr oder weniger Titeln und genauen Mengen- und Verhältnisangaben gefüllt ist; z.T. vielleicht nur mit drei oder vier Titeln großen Umfangs, z.T. mit mehreren hundert Titeln. Dieses Kastensystem kann man sich zeitlich gestuft denken, sofern man Querschnitte durch bestimmte Jahresproduktionen ermittelt, jedoch will ich die damit verbundenen Fragen hier übergehen.

2.4. Der nächste Schritt wäre die statistische Zwischenprüfung. Es ist wahrscheinlich, daß in den verschiedenen Gattungen [16] verschiedene statistische Strukturen herrschen. Es kann sein, daß man von bestimmten Gattungen weniger Text braucht, um eine Modellmenge zu erreichen, welche die statistischen Eigenheiten in einem gewünschten Grad widerspiegelt, von anderen Gattungen weit mehr. Hier wäre also das Bittmeiersche oder ein ähnliches Prüfverfahren einzusetzen, über das auf dieser Sitzung ja noch gesprochen werden wird [17]. Man würde dazu aus jedem der "Kästen" (Texttypen) eine eng umgrenzte Auswahl von Werken ent-

nehmen, aus diesen wieder eine höher konzentrierte, aber ebenfalls eng umgrenzte Auswahltextmenge, und ermittelt daraus den Wahrscheinlichkeitsquotienten, der ausdrückt, mit welcher Wahrscheinlichkeit man seltene Wortformen des Gesamttextes in der Auswahlmenge erwarten kann. Vermutlich wird sich zeigen, daß bei Zeitungen eine relativ große Textmenge erforderlich ist, um einen bestimmten Wahrscheinlichkeitsquotienten zu erreichen, bei Romanen eines bestimmten Typs weniger Text und etwa bei Trivialliteratur vom Typ Liebesroman wiederum weniger.
Auf Grund der an diesen Textproben ermittelten Wahrscheinlichkeitsverhältnisse innerhalb der einzelnen Untergattungen ist nun ein neues Kastensystem anzulegen, das entsprechend modifizierte Mengenverhältnisse aufweist.
Ein Ergebnis könnte beispielsweise sein, daß ich aus der Gattung Trivialliteratur-Untergattung Krimi, die in meinem Kasten mit 250 Titeln à 100 Seiten besetzt ist, nur 100 Seiten aufzunehmen brauche, um den gleichen Grad von Modellfähigkeit zu erreichen, wie bei 300 Seiten Text aus der Gattung technische Sachbücher, die mit 70 Titeln à 250 Seiten besetzt ist. D.h. ich erhalte zu allen Kästen bzw. Texttypen Mengenverhältnisse, die auf ihre statistische Modellfähigkeit hin modifiziert sind.

2.5. Auf Grund dieses neuen modifizierten Kastensystems kann man - mittels einer neuen Einschränkung - nun die Werke auswählen, die tatsächlich zur Berücksichtigung in den Modellmengen vorgesehen sind. Grundsätzlich muß bei dieser Auswahl der eigentlich triviale Satz gelten, daß statistisch gesehen 10 Prozent aus 100 Titeln immer besser sind als 100 Prozent aus 10 Titeln. Die Zahl der Titel kann also verhältnismäßig hoch gewählt werden.

Spätestens auf dieser Stufe muß aber nun eindeutig definiert werden, was für was als Modellmenge gelten soll.
Eine gutgestreute Auswahl, z.B. 5 Prozent aus 20 Titeln, ist an sich modellfähig nur für diese 20 Titel. Inwiefern diese 20 Titel wiederum aussagefähig,

exemplarisch für die Gattung oder Untergattung sind, aus der sie entnommen wurden, muß sorgfältig überlegt und begründbar sein, sofern nicht wieder objektive Kriterien wie Auflagenhöhe gewählt werden, was aber vielleicht nicht unbedingt erwünscht ist. Es kann sein, daß ein bestimmtes, an sich sehr auflagenschwaches Werk in besonderer Weise als exemplarisch für eine bestimmte Gattung oder Untergattung gelten kann. Im einzelnen können sich bei dieser Auswahl erhebliche theoretische und praktische Probleme ergeben, die zu lösen ich jetzt außerstande bin. Auch müßte geklärt werden, wie die bisher in Mannheim schon aufgenommenen Texte in das Corpus eingegliedert werden können, was selbstverständlich soweit wie möglich geschehen sollte. Eine spezielle kleine Arbeitsgruppe aus Fachleuten scheint mir zur Lösung all dieser Fragen erforderlich.

An dieser Stelle werden außerdem natürlich Gesichtspunkte der Kapazität wirksam. Diese können, falls unabänderlich, zu einschneidenden Verzichten zwingen. Wichtig ist jedoch, daß genau kalkulierbar bleibt, welche Auswirkungen bestimmte Verzichte auf die Widerspiegelung der Gesamt- und der Einzelverhältnisse sowie der Wahrscheinlichkeitsquote mit sich bringen.

Als letzter Schritt vor dem Beginn der Textaufnahme muß dann die Stichprobenverteilung in den zur Aufnahme vorgesehenen Werken geregelt werden. Was die Zeitungen betrifft, so liegen in der Außenstelle Bonn ja seit 1965 praktische Lösungsversuche für das Gesamtsystem der Textermittlung und -berechnung bis zur Herstellung der Modellmenge vor [18]. Ich möchte daher auf Einzelheiten hier verzichten. Grundsätzlich stellt sich die Stichprobenverteilung und die Ermittlung von Modellmengen bei Büchern als Vereinfachung des bei Zeitungen anwendbaren Verfahrens dar.

3. Zur Textaufnahme

Die Technik der Textaufnahme ist längst geläufig. Ein wesentlicher praktischer Unterschied gegenüber dem jetzigen Zustand würde der sein, daß wir es mit einer Vielzahl von Titeln zu tun hätten (vielleicht mit mehreren hundert) aus möglicherweise mehreren Dutzend Gattungen oder Texttypen; jeder Titel würde andererseits nur in einer Auswahl vorhanden sein.
Die Tatsache dieser sehr großen Zahl verschiedener Titel macht die Einführung eines Systems von Informationskarten praktisch unumgänglich, um die Einzeltexte bzw. die Titel, aus denen sie stammen, näher kennzeichnen zu können und sie somit überhaupt auf Abfrage verfügbar zu halten.

3.1. Dazu einige Vorschläge [19]:

Eine Informationskarte sollte selbstverständlich Verfassername, Werktitel und die genaue bibliographische Angabe enthalten, außerdem die Auflagenhöhe. Dann wären weitere Zahlenangaben erwünscht, etwa Umfang und Verteilung der Auswahl, Aufnahmequote, Wahrscheinlichkeitsquotient dieser Auswahl. Auch die Stellung des Titels innerhalb des Kastensystems sollte angegeben werden (etwa in Form eines Chiffrensystems analog zu dem von Kern vorgeschlagenen). Sodann wären weitere Angaben zum Titel denkbar, vielleicht eine kurze Angabe des Themas, dann literaturwissenschaftliche Kategorien (Gattung, Erscheinungsform, Stilhöhe, Gestaltungsweise u.ä.), vor allem auch intentionale Angaben. An dieser Stelle haben die von Kern vorgeschlagenen Kategorien unbedingt ihren sinnvollen Platz.
Sie müssen aber keineswegs die einzigen bleiben. Idealiter sollten so viele Informationskategorien berücksichtigt werden, wie einfache Fragestellungen denkbar sind, wobei sich natürlich durch Kombination mehrerer Informationskategorien zahlreiche weitere Fragestellungen an das Material richten lassen.

Damit ist der erste Schritt zur Dokumentation getan: wir verfügen über gegliedertes Material, das reproduzierbar ist unter möglichst vielen Fragestellungen. Jederzeit ist ein gezielter Zugriff, sind neue Ordnungs-, Umordnungs- und auch "Gewichtungs"möglichkeiten [20] gegeben.

Diese Informationskarten sind laufend den einzelnen Texten beizufügen, sobald diese abgeschrieben sind[21].

4. Zum Ausbau der Dokumentation

Nachdem damit nun eine Grundstufe der Dokumentation erreicht ist, folgt eine theoretisch und wohl auch praktisch nahezu unbegrenzte Zahl von Aufbau- und Ausbaustufen.

4.1. Bei dem weiteren Ausbau der Dokumentation ist nun zu berücksichtigen, in welchen Richtungen er vorgenommen werden soll. Wenn wir uns an die traditionelle Einteilung in Wortschatz, Morphologie und Syntax, vielleicht noch Stilistik, halten wollen, dann sind davon Wortschatz durch das "syntagmatische Arbeitsvorhaben" des Instituts für deutsche Sprache, Morphologie durch die bekannten Mannheimer Arbeitsvorhaben und ebenso Syntax in Mannheim direkt vertreten. Daraus ist die Folgerung zu ziehen, daß die Dokumentation in allen drei Richtungen ausgebaut werden müßte, oder anders gesagt: es ist ein Dokumentationssystem erforderlich, das zunächst den Texten solche Grundinformationen beifügt, die für alle drei Richtungen g e m e i n s a m relevant sind. Die Auswertung dieser Grundinformationen durch die drei Arbeitsgruppen oder auch durch auswärtige Wissenschaftler muß wieder neue, speziellere Informationen erbringen, die den Texten beizufügen sind. - Auf einer gewissen Stufe werden sich die Richtungen gabeln: speziellere Formen der dokumentierten Texte für die einzelnen Richtungen sind denkbar.

4.2. Schon im Ansatz ist also solche Dokumentation ein ständig wechselndes Mensch-Maschine-Verfahren; d.h. ein Wechsel von Informationseingabe, Informationsverarbeitung und -übertragung, Informationserzeugung (hier: maschinelle Textanalyse aufgrund gegebener Informationen mit dem Ziel der Erzeugung von Texten mit höherem Informationsgehalt) und Informationsausgabe (Textreproduktion bzw. Analyseausgabe) - dann folgt eine

neue Stufe der Aufarbeitung, der Formalisierung der Ergebnisse und der Informationseingabe.

Aber nicht nur der Übergang von einer Stufe zur höheren, sondern auch jeder einzelne Schritt innerhalb einer Stufe, wie etwa "Aufarbeitung" oder "Informationsübertragung", ist de facto wieder eine Folge mehrerer einzelner Mensch-Maschine-Schritte.

Diesem ganzen System muß ein entsprechender Programmierplan bzw. ein Programmiersystem parallelgeordnet sein [22]. Es ist mir unmöglich, hier Einzelheiten für ein solches System zu entwickeln. Dazu bin ich allein außerstande; es wäre außerdem zu kompliziert, um hier kurzerhand skizziert zu werden. Es handelt sich sicher um eine Aufgabe, die erhebliche Mühe guter Fachleute erfordert.

Eine Meinungsäußerung mag vielleicht möglich sein :

4.2.1. Ich glaube nicht, daß z.B. Wortschatzforschung von einer Dokumentation profitieren würde, die zunächst etwa auf syntaktische Fragestellungen ausgerichtet wird. Andererseits könnten aber sowohl Wortschatzforschung einschließlich Wortbildungsforschung als auch Morphologie und Syntax von Texten profitieren, in denen die Wörter (im Sinne lexikalischer Einheiten) mit Informationen ausgestattet sind. Ein Beispiel für die Ausstattung von Texten mit solchen Informationen ist die Rückführung der einzelnen Wörter auf ihre Grundformen unter Beigabe grober grammatischer Angaben. Ziel könnte ein automatisches kumulatives (und reversibles, d.h. in laufenden Text rückwandelbares) Wörterbuch sein, das dann nicht nur zu Wortschatzfragen Auskunft zu geben in der Lage ist. Vieles auf diesem - gewiß langen und schwierigen - Wege, wenn auch nicht alles, ist sicherlich formalisierbar. Ohne Zweifel müßten bei einem solchen Verfahren Wortschatz- und Grammatikexperten mit Programmierern eng zusammenarbeiten. Aber ein solches Verfahren, das im Bonner Institut für Phonetik und Kommunikationsforschung und in Saarbrücken schon in Angriff genommen worden ist und zu dem auch in Mannheim schon gewisse Vorarbeiten angelaufen sind, hat für das Mannheimer Institut nur Sinn, wenn es nicht

isoliert betrieben wird, sondern sinnvolle Stufe, sinnvolles Glied in einem Gesamtsystem ist.
Dokumentation einerseits und grammatische, lexikologische, syntaktische Arbeit andererseits müssen sich also von Stufe zu Stufe quasi gegenseitig "hinaufschaukeln". Ich bin davon überzeugt, daß sich für die einzelnen Arbeitsvorhaben und Wissenschaftler gerade dann bemerkenswerter praktischer Nutzen aus der Dokumentation ziehen läßt, wenn man grundsätzlich darauf verzichtet, ad hoc zu programmieren und zu "dokumentieren" (womit man den Anforderungen immer nur hoffnungslos nachläuft) oder Texte im Sinne einer Archivierungsaufgabe um ihrer selbst willen abzuschreiben, und sich statt dessen entschließt, von einer gemeinsamen Basis aus systematisch von Stufe zu Stufe ein sich allmählich differenzierendes Korpus hochinformierter Texte aufzubauen.

4.3. Dokumentation, wie ich sie verstehe, ist, wie jetzt wohl erkennbar, zwar einerseits durchaus - um frei nach Luther zu sprechen - "ein Knecht aller Arbeitsgruppen und jedermann untertan", ebenso aber ein eigenständiges Arbeitsprinzip und in gewisser Weise auch Mittelpunkt und Zentrum der verschiedenen Arbeitsrichtungen. Sie ist als zentraler Umschlagplatz vielfacher Informationen auf mehreren Informationsebenen für innen und außen zugleich eine ständige Aufgabe.
Nur dann scheint mir außerdem gewährleistet, daß nicht jeder Forscher bei der Textanalyse und -aufbereitung immer wieder von vorn anfangen muß, sondern daß einmal vorhandene Ergebnisse jederzeit und für jedermann zugänglich sind. Ich wüßte nicht, wer in der Bundesrepublik sich einer solchen Aufgabe jemals unterziehen könnte, wenn nicht das Institut für deutsche Sprache.

5. Zur Organisation

Ein solches Verfahren verlangt allerdings ein Neudurchdenken der Prioritäten und ein Denken in längeren Zeiträumen. Und vor allem wird es erhebliche organisatorische Änderungen und Umgruppierungen verlangen. Dokumentation ist ja nicht einfach Ergebnis der Arbeit unserer Schreibdamen und der Korrektoren. Dokumentation ist auch nicht Sache von Programmierern oder Sache der Wortschatz- und Grammatikspezialisten als solcher. Dokumentation ist ein Gebiet eigener Verantwortlichkeit. Unerläßlich sind allerdings Programmierer; unerläßlich ist auch der Grammatikexperte, der in der Lage ist, wissenschaftliche Ergebnisse im Bereich der Morphologie und der Syntax zu formulieren bzw. geeignete Fragen an die Texte der betreffenden Informationsstufe zu richten. Unerläßlich wird auch ein Wortschatzexperte sein sowie ein weiterer Mitarbeiter, den wir vielleicht Corpus-Statistiker nennen können, der die Aufgabe hätte, laufend die zur Aufnahme vorgesehenen Texte zu ermitteln. Dokumentation würde l'art pour l'art, wenn sie die Mitarbeit und Zusammenarbeit dieser Experten entbehren müßte. Kurz gesagt: sie bietet Gelegenheit zum teamwork, wo es am dringendsten nötig ist.

Folgende personelle Gliederung und Zusammensetzung schiene mir zur Erfüllung der Dokumentationsaufgaben sinnvoll und erforderlich: ein selbständige Abteilung mit fünf bis sechs wissenschaftlichen Mitarbeitern, vier bis sechs ganztägigen Schreibkräften und ebenso viel Korrektoren sowie zwei bis drei studentischen Hilfskräften. Zur Unterstützung dieser Arbeitsgruppe wäre ein kleines, leicht erreichbares Gremium von je einem Fachmann aus den Bereichen Statistik, Dokumentation, Datenverarbeitung, Grammatik und Literaturwissenschaft erwünscht und wohl auch notwendig.

Durch eine solche Abteilung würde vermutlich die geplante eigene Rechenanlage voll ausgelastet; sie wäre in der Lage, Bedürfnisse der Mitarbeiter besser als bisher zu befriedigen, Anforderungen von außerhalb auch auf längere Sicht voll zu erfüllen und vor allem eine Dokumentation zu betreiben, die diesen Namen verdient und nicht nur der Mannheimer Linguistik, sondern generell der Linguistik,

soweit interessiert, auf lange Sicht etwas nützt.

Herr Professor von Polenz wird es mir hoffentlich nicht verübeln, wenn ich zum Schluß den letzten Satz seines Vortrages vom Frühjahr 1965 leicht abgewandelt zitiere: "Ich darf also mit der Hoffnung schließen, daß sich das Institut für deutsche Sprache schon in der Gesamtkonzeption seiner Dokumentation solcher Aufgaben annimmt" [23].

Anhang

Zusammenfassung des Mannheimer Corpus[24]

von Manfred W. Hellmann

Die folgende Übersicht erstrebt weder eine geschlossene Systematik noch Vollständigkeit der möglichen Gesichtspunkte. Sie will einen Überblick über das Vorhandene ermöglichen und zugleich Hinweise auf denkbare Gliederungsschemata geben.

Grundlage dieser Untersuchung ist die Quellenliste des Mannheimer Corpus in der Fassung vom Oktober 1967 [25]. Ich bitte, die in dieser Liste aufgeführten Titel sich fortlaufend durchnumeriert zu denken (Nr. 1 = Bergengruen, Tempelchen; Nr. 27 = "Bild Zeitung").

Ein Überblick über die Quellenliste vermittelt den Eindruck eines Zuges zur literarischen Qualität. Ausnahmen bilden nur die wenigen Seiten der Gruppe b (Trivialliteratur) und die mengenmäßig allerdings bedeutsame Auswahl aus der "Bild Zeitung".
Der Aufbau der Quellenliste verrät die Wirksamkeit zweier verschiedener Kategorien: während man die Begriffe "Dichtung", "Trivialliteratur", "wissenschaftliche und populärwissenschaftliche Literatur", "Memoiren" als Gattungsbegriffe bezeichnen kann, sind mit den Begriffen "Zeitungen", "Zeitschriften" (Gruppe e) Erscheinungsformen (Druckformen, Veröffentlichungsformen) angesprochen.

1. Zur Verteilung nach "Gattungen" (Intentionen)[26]

Bei den Gruppen a (Dichtung) und b (Trivialliteratur) scheinen sich intentionale Kategorien mit solchen der literarischen Qualität zu verbinden. Gemeinsam ist ihnen, daß es sich um "erzählende" Literatur handelt ("Fiction", "Belletristik"); aber während dabei die Titel der Gruppe a durchweg der höchsten Intention in dieser Gattung folgen, für die ich den Begriff "Künstlerische Weltgestaltung" verwenden möchte, folgt die Gruppe b der untersten Intention der "Unterhaltung". Die Mittellage "Gehobene Unterhaltung" fehlt.

Die Gruppen c (Wissensch. u. populärwiss. Literatur) und d (Memoiren) könnte man der Intention "Information, Unterrichtung" zuordnen, deren höchste Form, z.B. in wissenschaftlichen Werken hohen Ranges, "Erkenntnisvermittlung" genannt werden könnte. Ausnahme in dieser Gruppe ist Nr. 20 als (juristische) Anleitung Belehrung. Andererseits gehören die Texte der populärwiss. Zeitschriften sicher in die Rubrik "Information, Unterrichtung".

Normalerweise finden sich in Zeitungen alle Intentionen des Schreibens nebeneinand Bei der FAZ und der WELT handelt es sich jedoch ausschließlich um Nachrichtentexte von der jeweils ersten Seite, die ohne weiteres ebenfalls der genannten Rubrik zuzuordnen sind. Nur die Texte der Bild Zeitung stellen einen Querschnitt dar[27]. Die Gattung "Anleitung, Unterweisung, Belehrung" ist demgegenüber mit dem einen juristischen Beraterbüchlein unterrepräsentiert.
Gleiches gilt für die intentionale Gattung "Werbung, Aufforderung", die nur durch einige "Bild"-Texte in unbekanntem Umfang vertreten ist.

II. Zur Verteilung nach Erscheinungsformen.

Die gröbste Einteilung ist die nach einmalig bzw. periodisch erscheinenden Druckerzeugnissen (Bücher - Periodika), dazu kommen die meist vernachlässigten Formen der Gelegenheitserzeugnisse wie Drucksachen, Flugblätter, Plakate u. dergl.
Unter der Erscheinungsform "Buch" sind Untergliederungen denkbar, die aber nicht relevant zu werden brauchen[28]. Einen Übergang zur Erscheinungsform "Periodika" stellen die broschierten Reihen ("Broschüren") dar. Periodika sind nur mit den Untergruppen "Tageszeitungen" und (unterrichtende) "Zeitschriften" vertreten; verschiedene sehr verbreitete Typen wie "Illustrierte" fehlen. Ebenso fehlt die letzte Grupp der gelegentlich erscheinenden "Drucksachen" völlig.

III. Zur Verteilung nach Sachgebieten.

Die bei weitem größte Gruppe stellt hier die nicht sachgebietsgebundene Literatur eine Folge der sehr starken Aufnahme von Büchern der Gattung "Dichtung" (Künstlerische Weltgestaltung). Die Sachgebiete Natur- und Geisteswissenschaften sind -

wenigstens mit bestimmten Untersachgebieten - noch relativ reichlich vertreten. Andere, viel gelesene, wie Sport, Technik, Soziales, Wirtschaft sind nur oder fast nur durch Texte in unbekannter Menge aus der Bild Zeitung vertreten und damit sicher unterrepräsentiert. Über die Verteilung der Sachgebiete in den populärwissenschaftlichen Zeitschriften lagen mir keine Angaben vor.
Im übrigen ist diese Liste der Sachgebiete sehr grob und pragmatisch und orientiert sich am Vorhandenen[29]; eine theoretisch - systematische Aufstellung würde Lücken und Ballungen sehr viel stärker sichtbar machen, aber auch den Überblick erschweren.

Die folgenden tabellarischen Übersichten beruhen auf den Seitenzählungen der Titel selbst, enthalten daher aufgrund des z.T. stark unterschiedlichen Seitenumfangs einen Unsicherheitsfaktor. Die Zeitungsseiten wurden 1 : 6 umgerechnet (1 Zeitungsseite = 6 Buchseiten).

Mengenmäßige Verteilung des Mannheimer Corpus

I. Verteilung des Corpus auf "Gattungen"

I. Belletristik			Zahl der Seiten
a) Dichtung (Künstlerische Weltgestaltung)			
(1)	Bergengruen	Tempelchen	41
(2)	Böll	Ansichten e. Clowns	292
(3)	Frisch	Homo Faber	245
(4)	Grass	Blechtrommel	703
(5)	Mann	Die Betrogene	122
(6)	Strittmatter	Ole Bienkopp	358
			1761
b) Gehobene Unterhaltung(sromane)			
(-)	- - -	- - -	-
c) Trivialliteratur			
(Heimat- und Liebesromane):			
(7)	Jung	Magd v. Zellerhof	59
(9)	Stauffen	Solange dein Herz schlägt	61
(Krimi:)			
(8)	Pinkwart	Mord ist nichts . . .	175
			295
(Science Fiction:)		- - -	-
(Kriegsliteratur)		- - -	-
(Erotische Literatur)		- - -	-
(Abenteuer u. Wildwest)		-	
Dazu: Geschätzter Anteil aus (27) BILD: ca.			130
			425
		Summe zu 1:	2176
			~2200

2. Informierende (unterrichtende) Literatur			Zahl der Seiten	
a) Wissenschaftl. u. populärwiss. Literatur				
(10)	Bamm	Ex Ovo	256	
(11)	Bollnow	Maß und Vermessenheit	232	
(12)	Gail	Weltraumfahrt	137	
(13)	Grzimek	Serengeti	325	
(14)	Heimpel	Kapitulation v.d. Geschichte	113	
(15)	Heisenberg	Naturbild d. Physik	39	
(16)	Jaspers	Atombombe	500	
(17)	Jungk	Zukunft hat schon begonnen	312	
(18)	Pörtner	Erben Roms	480	
(19)	Staiger	Grundbegr.d. Poetik	249	
			~2640	
(24)	Zs.	Bild d. Wissenschaft (3 H.)	360	
(25)	Zs.	Studium Generale (1 H.)	120	
(26)	Zs.	Urania (2 H.)	150	
			630	
			3270	
b) Memoiren				
(21)	Heuss	Erinnerungen	439	
c) Aktuelle Information				
(22)	Tagesztg.	FAZ, 25 mal 1. Seite (1 Zeitgsseite = 6 Seiten)	150	
(23)	Tagesztg.	DIE WELT, ca 65 mal 1. Seite	390	
(24)	Tagesztg.	BILD (Querschnitt durch 7 Monate, Tgl. 1 Seite = ca. 175 Seiten, davon 2/3 (mal 6) : ca.	720	
		Summe c:	1260	
		Summe b:	439	
		Summe a:	3270	
		Summe zu 2:	4969	4969
				~5000

3. Anleitende (belehrende) Literatur (Gebrauchsliteratur)			
a) Fachliche Anleitung			
(20)	Ullrich	Wehr dich Bürger! (Jur.)	153
b) Gebrauchsanweisungen			--

	Zahl der Seiten	
	153	
c) Kochbücher	-- --	
d) Gesundheits-Hausbücher	--	
e) Techn. o. Bastelbücher	--	
f) Reisebücher (-führer)	--	
g) Sonstige	--	
	153	153

4. Werbende u. auffordernde Literatur

a) Kommerzielle Werbung		
1. Werbekataloge	--	
2. Werbeprospekte	--	
3. Werbeanzeigen	?	
b) Politische Werbung		
1. Politische Aufrufe (Programme)	--	
2. Politische Anzeigen	?	
c) Soziale u. weltanschauliche Werbung		
1. Caritative Aufrufe (Rundschreiben)	--	
2. Caritative Anzeigen	?	
3. usw.		
Zu 4 insgesamt: unklarer Anteil aus (27) BILD ca.	200	200

Summe der Seiten

(bei Zeitungen und Zeitschriften geschätzt):

aus 1 : 2200
aus 2 : 5000
aus 3 : 153
aus 4 : 200

7553
~ 7500

II. Verteilung des Corpus auf "Erscheinungsformen"

		Zahl der Seiten	
1. Bücher			
a) mit einheitlichem Inhalt	Nr. 1 - 6	1761	
	Nr. 10 - 19	2640	
	Nr. 21	439	
		4840	
b) mit verschiedenem Inhalt (Sammelwerke)		--	
			4840
2. Broschüren			
	Nr. 8	175	
	Nr. 20	153	
		328	
			328
3. Periodika			
a) Zeitungen			
1) Tageszeitungen:			
überregional	Nr. 22 - 23	540	
regional			
Boulevard	Nr. 27	1050	
		1590	
2) Wochenzeitungen:		--	
seriös		--	
unseriös		--	
b) Zeitschriften			
1) Unterhaltungs Zs. (Illustrierte)		--	
2) informierende Zs. (Magazine)		--	
3) Speziellere Zs.			
wissenschaftliche		--	
populärwissensch.		630	
kulturelle		--	
literarische		--	
politische		--	
usw.			
c) "Hefte" (in Reihen)			
Heimat- u. Liebes-	Nr. 7 - 8	120	
Krimis (als Reihen-Heft!)			
Science Fiction		--	
Kriegs-		--	
Erotica		--	
Abenteuer-Wildwest			
Sonstige		--	
		750	
			2340

4. Drucksachen

a) kommerzielle

Werbeprospekte, Gebrauchsanw.
Angebote usw.

b) öffentliche

Wahlpropaganda
Bekanntmachungen
Aufforderungen
usw.

c) private

Kinoprogramme
Flugblätter
usw.

entfällt

5. Plakate

a) kommerzielle

b) öffentliche entfällt

Summe aus 1 :	4840
Summe aus 2 :	328
Summe aus 3 :	2340
Summe aus 4 :	--
Summe aus 5 :	--
	7508

III. Verteilung des Corpus auf "Sachgebiete"

			Zahl der Seiten
0. Nicht sachgebietsgebunden (Fiction):	Nr. 1 - 6 und 7 - 9		2056
1. Politik (Außen-, Innen- usw.)	Nr. 21, Nr. 22, 23 (Nr. 27)	ca.	1250
2. Wirtschaft Industrie Handel Finanzen Landwirtsch. usw.	Nr. 27	ca.	200 (?)
3. Technik	Nr. 26? Nr. 12	=	287
4. Kultur	Nr. 10, 18 (+25?)	=	736 (+ 120)
5. Kunst	nicht vorh.		
6. Soziales	(Nr. 27)		150 (?)
7. Naturwissenschaft Physik Chemie Biologie (Zoologie) Medizin usw.	Nr. 15 Nr. 13 Nr. 17, 24	zus.	1030
8. Geisteswissenschaft Rechtswiss. Literaturwiss. Theologie Philosophie Geschichte usw.	Nr. 20 Nr. 19 Nr. 11, 16 Nr. 14	zus.	1245
9. Sport	(Nr. 27)	ca.	150
10. Sonstiges Naturereignisse, Wetter Verbrechen Werbung usw.	Nr. 23, 27? Nr. 27 ? Nr. 27 ?	ca.	300 ?
		ca.	7500

Anmerkungen

1) Es handelt sich hier um die leicht geänderte und erweiterte Wiedergabe meines am 29. Juni 1968 auf der Sitzung der Kommissionen für Dokumentation und für datenverarbeitende Maschinen vorgetragenen Diskussionsbeitrages. Die Einleitung wurde neu formuliert.

2) Vor allem durch Peter von Polenz, "Zur Quellenwahl für Dokumentation und Erforschung der deutschen Sprache der Gegenwart" in : Wirkendes Wort 16, 1966, S. 3-13, und in : Satz und Wort im heutigen Deutsch. Sprache der Gegenwart Bd. 1. Schriften des Instituts für deutsche Sprache in Mannheim, Düsseldorf 1967, S. 363-378 (ursprünglich Vortrag auf der Frühjahrstagung des Mannheimer Instituts 1965). Zitiert wird im folgenden nach dem Druck in "Sprache der Gegenwart".

3) Eine Liste des Mannheimer Corpus (allerdings ohne Mengenangaben) findet sich im Forschungsbericht 2 (1968) des Instituts für deutsche Sprache Mannheim (Rotaprintdruck), Teil I, Seite 11; das Verfahren der Textaufnahme ist im selben Forschungsbericht dargelegt.

4) Vgl. Verf., Dokumentation und maschinelle Textverarbeitung in der Außenstelle Bonn. Forschungsbericht 2 (1968) des Instituts für deutsche Sprache, Teil II (Seite 39 ff.).

5) Peter von Polenz, a.a.O. S. 363.

6) ebd. S. 366.

7) ebd. S. 368.

8) ebd. S. 369.

9) vgl. ebd. S. 373.

10) Ein Teil der Mannheimer Mitarbeiter hat Erfahrungen gemacht, die diese Unterlegenheit einer so verstandenen "Dokumentation" klar demonstrieren. Allerdings litt die Mannheimer Textverarbeitung auch unter Schwierigkeiten, die vom Institut nicht zu vertreten sind.

11) Es wäre sicherlich möglich, durch eine Art betriebswirtschaftlicher Kostenrechnung eine Kostenanalyse zu erstellen, um wieviel Prozent sich die Kosten relativ zum Nutzen erhöhen, wenn bestimmte Einschränkungen vorgenommen werden, und wann der Punkt erreicht ist, bei dem voraussichtlich die Kosten den auf Dauer (20 Jahre) erzielbaren Nutzen übersteigen.

12) Systeme wie das von Kern vorgeschlagene scheinen mir daher für den ersten Schritt nicht geeignet, weil sich die einzelnen Einheiten solcher Systeme nicht quantifizieren lassen, ohne daß man die für jede Einheit infragekommenden Titel kennt. Eine solche Kenntnis ist aber auf dieser Stufe ersichtlich nicht gegeben.

13) Vor allem natürlich die Verlage selbst könnten Auskunft geben; vielleicht auch das Institut für Buchmarktforschung und einige Universitätsinstitute. Auch die Presse bringt gelegentlich solche Statistiken.

14) Dagegen scheint es mir unmöglich, etwa Zahl und Umfang derjenigen Schriften festzustellen, auf welche etwa die Kernschen Kriterien "Du-Orientierung - Überredung" zutreffen, oder auch die einem Texttyp wie "Leitartikel - Kommentar" zugehörigen Texte.

15) Die Angaben beziehen sich auf den Stand von September 1967. Vgl. Walter J. Schütz, Veränderungen im deutschen Zeitungswesen zwischen 1954 und 1967. In: Publizistik 12. Jg. 1967, H.4, S. 243-246. - Der Begriff "Tageszeitung" umfaßt hier auch Ausgaben von Zeitungen, die sich nur teilweise, z.B. im Lokalteil, von einander unterscheiden.

15a) Die Angaben über den durchschnittlichen Seitenumfang verdanke ich einer freundlichen Mitteilung von Herrn K.H. Teckentrup, Mainz. Er hat in einer noch nicht veröffentlichten Auszählung für 224 der größeren deutschen Zeitungen im Jahre 1967 einen Durchschnittsumfang von 24 Seiten (auf mittleres Zeitungsformat umgerechnet) ermittelt. Bezieht man die zahlreichen kleineren Zeitungen in die Berechnung mit ein, ergibt sich ein Durchschnittswert von nicht weit unter 20.

16) "Gattungen" hier verstanden als Summe der Werkmengen (Titelmengen) des betreffenden "Kastens".

17) Das am 29. Juni gehaltene Referat von Günther Billmeier stellt eine Zusammenfassung seines Aufsatzes "Über die Signifikanz von Auswahltexten" dar, der im Forschungsbericht 2 (1968) des Instituts für deutsche Sprache, Teil III (Seite 126 ff.), erschienen ist.

18) Siehe Anmerkung 4.

19) Vgl. auch das (allerdings zur Kennzeichnung von Zeitungstexten) in der Außenstelle Bonn entwickelte System von Informationskarten (Forschungsbericht 2, Teil II, Anhang III).

20) Gewichtungen sind immer eine Art von Vorinterpretation. Sie sind daher meines Erachtens bei der Ermittlung der zum Corpus gehörigen Texte und der Festlegung ihrer Mengenverhältnisse tunlichst aus dem Spiel zu lassen. Sie sollten unbedingt dem einzelnen Forscher im Rahmen seiner speziellen Fragestellung beim Umgang mit den (einmal vorhandenen) Corpustexten vorbehalten bleiben; hier sind sie sogar erforderlich. Als Beispiel für eine mir sinnvoll scheinende Gewichtung aus der Arbeit mit Zeitungstexten sei folgendes erwähnt: Gesucht werden idiomatische Ausdrücke wertender Art; sie sollen u.a. auf ihren Propagandawert untersucht werden. Bekanntlich hat die Seite 1 bei allen Zeitungen einen höheren Propagandawert als die übrigen Seiten, die Seite 2 einen niedrigeren als Seite 1; die letzte Seite dürfte einen zwischen 1 und 2 liegenden

Propagandawert besitzen. Überschriften haben wiederum einen höheren Propagandawert als normaler Text. Eine entsprechende Gewichtung könnte zum Ergebnis haben, daß alle Belege von vorn herein in einer auf meine Frage zugeschnittenen Perspektive erscheinen: sie sind vorinterpretiert auf schnellste und einfachste Weise. Ebenso leicht läßt sich eine solche Gewichtung wieder tilgen, um Platz für andere Fragestellungen und damit andere Gewichtungen zu machen.

21) Es braucht wohl kaum erwähnt zu werden, daß zu diesem Zeitpunkt ein Programm zur Auswertung der Informationskarten und Ausgabe der in ihnen enthaltenen Informationen zusammen mit Kontext-Beleglisten oder Registern vorhanden sein muß. Solche Programme haben sich als für jede Art von Arbeit an Texten nützlich erwiesen.

22) In der Außenstelle Bonn ist im Dezember 1966, zusammen mit dem Institut für Phonetik und Kommunikationsforschung, einmal der Versuch gemacht worden, ein Programmiersystem zur Erstellung eines kumulativen Wörterbuchs (allerdings überwiegend für Wortschatzfragen) zu entwerfen.

23) Peter von Polenz, a.a.O. S. 378.

Anmerkungen zum Anhang

24) Der "Anhang" ist eine Teil-Zusammenfassung eines Diskussionsbeitrages auf der Sitzung der Kommission für Dokumentation am 29. Juni 1968. Ziel des Beitrages war es, neben einer Analyse des vorhandenen Corpus konkrete Vorschläge zu seiner gezielten, begrenzten Erweiterung vorzulegen. Die Vorliegende Zusammenfassung beschäftigt sich jedoch nur mit der Analyse.

25) Vgl. jetzt Forschungsbericht 2 des Instituts, S. 11 f.

26) Bei der Einteilung nach intentional bestimmten Gattungen lehne ich mich an die vornehmlich in der Publizistik geläufige Gliederung nach Schreibintentionen an:

Unterrichtung, Belehrung, Beeinflussung, Unterhaltung und Werbung. Es handelt sich hier um einen Vorschlag, dessen Anwendbarkeit auf ein allgemeineres Corpus noch geprüft werden müßte; sicherlich sind Änderungen, z.B. bessere Untergliederungen, möglich und nötig.

27) Über die Zusammensetzung des Materials aus der Bild Zeitung lagen mir keine Angaben vor.

28) Denkbar wäre eine Einteilung nach der Einheitlichkeit des Inhalts.

29) Zu dem in der Außenstelle Bonn für Zeitungstexte verwendeten, ebenfalls pragmatischen, Schema von "Sachgebieten" vgl. Forschungsbericht 2, S. 99-102.

Teilerhebungen und ihre Anwendung auf die Sprachbearbeitung

von Werner Müller

Übersicht:

1. Allgemeines zu Teilerhebungen (56)
   1.1. Stichprobenerhebung und Repräsentativerhebung (56)
   1.2. Vorteile der Teilerhebung (57)

2. Zum Stichprobenverfahren (58)
   2.1. Allgemeines zu den Stichproben (58)
   2.2. Darstellung und Anwendung von Stichprobenverfahren (59)
      2.2.1. Uneingeschränkte Zufallsauswahl heterograder Natur (59)
      2.2.2. Uneingeschränkte Zufallsauswahl homograder Natur (61)
      2.2.3. Uneingeschränkte Zufallsauswahl bei sehr großem oder unbekannt großem Umfang der Population (63)
      2.2.4. Eingeschränkte Zufallsauswahl im Gegensatz zu uneingeschränkter Zufallsauswahl (63)
      2.2.5. Die geschichtete Stichprobe (64)
      2.2.6. Die geschichtete Stichprobe mit prozentualer Aufteilung (67)

3. Allgemeines zur Anwendung der geschichteten Stichprobe bei Corpusuntersuchungen (71)

Anmerkungen (73)

Literatur (74)

Bemerkung:

Diese Studie enthält weder methodisch grundsätzlich Neues noch Ergebnisse von Untersuchungen am Textcorpus des Instituts für deutsche Sprache. Vielmehr soll gezeigt werden, wie statistische Methoden, wenn man sie in der praktischen Sprachbearbeitung anwendet, auf einfachere Art zu zuverlässigen Ergebnissen führen können.

Einleitend:

Ziel vieler Untersuchungen des Instituts für deutsche Sprache ist es, Zahlen und zahlenmäßige Gefüge zu ermitteln, die für die Sprache allgemeine Gültigkeit besitzen. Eine so geartete Zielsetzung bedarf der Anwendung statistischer Methoden, ist doch die Statistik notwendige Hilfswissenschaft zur numerischen Untersuchung von Massenerscheinungen. Aussagen über Massenerscheinungen werden auf Grund von Teilerhebungen gemacht, wobei die Teilerhebungen einzuteilen sind in Repräsentativ- und Stichprobenerhebungen.
In diesen Ausführungen sei die Anwendung von Methoden der Teilerhebung, besonders der Stichprobenerhebung, auf die Sprachbearbeitung beschrieben, wozu es notwendig erscheint, Stichprobenmethoden unter Anführung von Beispielen zu erläutern[1]. Dann soll beschrieben werden, wie man mittels dieser Methoden zu zahlenmäßigen Aussagen gelangt, die als gültig für die Sprache erachtet werden können[2].

1. Allgemeines zu Teilerhebungen

1.1. Stichprobenerhebung und Repräsentativerhebung

Teilerhebung bedeutet ganz allgemein, daß nicht alle Untersuchungselemente der Grundgesamtheit einer Untersuchung unterzogen werden, sondern nur ein Teil dieser Menge. Wie gelangt man nun von einer Grundgesamtheit zu einer Untersuchungsmenge? Einerseits kann man die Untersuchungselemente nach Gesichtspunkten der Repräsentativität aus der Grundgesamtheit wählen, andererseits kann man sie auch durch den Zufall herausgreifen lassen. Ersteres wird als Repräsentativerhebung, letzteres als Stichprobenerhebung bezeichnet. Jede Art von Teilerhebung kann nur durchgeführt werden, wenn zwei notwendige Voraussetzungen erfüllt sind: Die Grundgesamtheit, die 'Population', muß nach Ort, Zeit und Sache genau bestimmt und abgegrenzt sein; die zu untersuchenden Einheiten müssen einwandfrei identifizierbar und auswertbar sein. Ein Beispiel einer abgegrenzten Grundgesamtheit: Die in der Bundesrepublik veröffentlichte

populärwissenschaftliche Literatur des Jahres 1968; die Formen des finiten Verbs.

Plant man eine Stichprobe, so muß zunächst ein Auswahlplan festgelegt werden, der garantiert, daß jedes Untersuchungselement der Population durch das Wirken des Zufalls die Chance besitzt, in die Stichprobe zu gelangen. Wenn der Auswahlplan dieser Forderung genügt, gestattet es die mathematische Statistik, von Gesetzmäßigkeiten der Auswahlmasse auf entsprechende Gesetzmäßigkeiten der Grundgesamtheit zu schließen.

Parameter der Grundgesamtheit bestimmen sich nach analogen Parametern der Erhebungsmasse unter Einbeziehung einer bestimmbaren Genauigkeit und einer bestimmbaren Sicherheit. Der Sicherheitsgrad, der den Aussagen über die Grundgesamtheit beigemessen wird, richtet sich nach der Wahrscheinlichkeit dafür, daß die erhobene Stichprobe eine Verallgemeinerung ihrer Ergebnisse gestattet. Bei der Verallgemeinerung des Untersuchungsergebnisses wird für das Stichprobenergebnis ein Vertrauensintervall berechnet, innerhalb dessen es für die Population gültig ist. In diese Berechnung geht der Sicherheitsgrad ein. Wir können Stichprobenverfahren demnach definieren als "Teilerhebungen, bei denen der Fehler, der durch Beschränkung auf einen Teil der Gesamtheit entsteht, berechenbar ist. Dies deshalb, weil dann die Wahrscheinlichkeitsrechnung und die mathematische Statistik angewandt werden können"[3].

1.2. Vorteile der Teilerhebung

Fragen wir uns, warum wir Teilerhebungen durchführen sollen. Gegenüber einer Totalerhebung hat die Teilerhebung den Vorteil, daß sie meist schneller zu Ergebnissen führt, daß die Erhebung meist von geringerem Kostenaufwand ist, daß die Erhebung wegen der geringeren Anzahl von Untersuchungselementen gründlicher durchgeführt werden kann. Freilich kann - wie im Fall der Sprache als Population - eine Totalerhebung von vornherein unmöglich sein; dann muß eine Teilerhebung a priori in Betracht gezogen werden. Die stichprobenartige Erhebung hat gegenüber der repräsentativen Erhebung den großen Vorteil, daß sie, wie oben schon gesagt, eine Fehlerberechnung zuläßt.

2. Zum Stichprobenverfahren

2.1. Allgemeines zu den Stichproben

Grundsätzlich sind alle möglichen Arten von Stichproben einteilbar nach folgenden zwei Unterscheidungskriterien : Liegt ihnen eine homograde oder heterograde Fragestellung zu Grund? Stellen sie eine uneingeschränkte oder eine eingeschränkte Zufallsauswahl dar?

Uneingeschränkte Zufallsauswahlen sind dadurch gekennzeichnet, daß jedes Untersuchungselement der Population die genau gleiche Chance besitzt, in die Stichprobenmasse zu gelangen, wohingegen bei eingeschränkten Zufallsauswahlen nur gesagt werden kann, daß jedes Element eine Chance besitzen muß, in die Erhebungsmasse zu gelangen, eine Chance, die von der eines anderen Untersuchungselements verschieden sein kann. Zum Beispiel : Ich denke mir als Grundgesamtheit die Masse aller bestimmt definierten Untersuchungselemente von Werk 1 und Werk 2. Will ich etwa 10% dieser Population nach einer uneingeschränkten Zufallsauswahl stichprobenartig untersuchen, so lasse ich den Zufall die einzelnen Untersuchungselemente aus der Population greifen. Diese uneingeschränkte Zufallsauswahl erfährt eine Einschränkung, wenn man - um bei dem angeführten Beispiel zu bleiben - wiederum 10% der Population erheben will, jedoch die Stichprobenerhebung so anordnet, daß sich in ihr 80% Untersuchungselemente von Werk 1 und 20% Untersuchungseinheiten von Werk 2 befinden. Zwar bleibt jedem Element der Population eine Chance, in die Stichprobe zu gelangen, doch ist diese Chance nicht mehr gleich für jedes Element der Grundgesamtheit, denn Untersuchungselemente von Werk 1 besitzen eine größere Chance, in die Stichprobe zu gelangen, als solche von Werk 2.

Ob eine Stichprobe homograder oder heterograder Natur ist, hängt von der zugrunde liegenden Fragestellung ab. Wird nach einem qualitativen Merkmal gefragt (zum Beispiel: Steht ein Verbum im Passiv? oder : Welcher Wortklasse gehört ein Wort an? so ist die Stichprobe homograder Natur. Ergebnisse von Stichproben homograder Natur werden in Anteilswerten wiedergegeben, zum Beispiel : 7,52% der Finita sind Konjunktive. Richtet sich die Fragestellung dagegen auf quantitive Merkmale (zum Beispiel: Wieviele Wörter stehen in einem Satz? oder : Welcher Abstand besteht von Substantiv zu Substantiv?), so ist die Stichprobe heterograder Natur. Ergebnisse quantitati

Fragestellungen werden in Durchschnittswerten angegeben, zum Beispiel: "Die durchschnittliche Satzlänge in TEMP[4)] beträgt 19,598 Tokens".

2. Darstellung und Anwendung von Stichprobenverfahren

2.2.1. Uneingeschränkte Zufallsauswahl heterograder Natur

Als Beispiel für eine Stichprobe heterograder Natur unter Zugrundelegung einer uneingeschränkten Zufallsauswahl sei die Bestimmung der durchschnittlichen Satzlänge von TEMP angeführt.

Es sei die Aufgabe gestellt, die durchschnittliche Satzlänge von TEMP innerhalb eines Vertrauensintervalls von + 1,5 Tokens zu bestimmen, wobei das Ergebnis einen Sicherheitsgrad von 95% haben soll.

Das TEMP enthält 431 Sätze. Zunächst gilt es, ein Zufallswahlverfahren zu bestimmen, das gewährleistet, daß wirklich jeder einzelne Satz von TEMP in die Stichprobe gelangen kann und daß jeder Satz hierfür die gleiche Chance von 1/431 besitze. Ein mögliches Auswahlverfahren, das diesen Bedingungen genügt, kann gegeben sein, wenn man sich vornimmt, nach einer Tabelle mit Zufallszahlen aus den durchnumerierten Sätzen einzelne Sätze herauszugreifen. Nun erhebt sich sofort die Frage, wieviele Sätze überhaupt zufällig herausgegriffen werden müssen. Dies ist die Frage nach dem Stichprobenumfang 'n'. Den für unsere Aufgabenstellung nötigen Stichprobenumfang kann man errechnen. Generell bedarf es bei der Abschätzung des notwendigen Stichprobenumfanges einerseits der Fehlergrenze und des Sicherheitsgrades; dies ist in dieser Aufgabenstellung fixiert. Andererseits bedarf es eines ungefähren Überblicks darüber, wie stark die einzelnen Messwerte für jedes Untersuchungselement um den wahren Mittelwert streuen. Hierzu bedarf es einer Voruntersuchung, bei der zunächst einmal 'n^*' Sätze stichprobenartig untersucht werden müssen. Aus den Ergebnissen der Voruntersuchung läßt sich ein Mittelwert $\bar{x}^*$ und die Varianz s^{*2} berechnen. Für die Varianz der Voruntersuchungsergebnisse gilt:

$$s^{*2} = \frac{\sum (x_i - \bar{x}^*)^2}{n^*} \qquad (1)$$

Der Wert für die Varianz der Voruntersuchungsergebnisse wird als mutmaßlicher Wert für die Varianz der Stichprobenergebnisse genommen. Bei einer Untersuchung von 25 zufällig gewählten Sätzen des TEMP ergibt sich nach (1) ein Wert für s^{*2} von 167. Es ist folglich anzunehmen, daß die wahre Varianz wohl kaum über 200 liegen wird. Mit diesem Wert für s^{*2} nebst den Werten, die festgesetzt wurden, kann man nun den notwendigen Stichprobenumfang rechnerisch abschätzen nach

$$n \geq \frac{t^2 \cdot N \cdot s^{*2}}{t^2 \cdot s^{*2} + (N-1) \cdot e^2} \qquad (2)$$

In unserem Beispiel ist N, der Umfang der Grundgesamtheit, N = 431 Sätze; s^{*2} wi ermittelt, man kann eine gewisse Sicherheitsspanne berücksichtigen, indem man etwa hier $s^{*2} = 200$ setzt; e war in der Aufgabenstellung gesetzt = 1,5; und einem Sicher heitsgrad von 95% entspricht ein t = 1,96 . So errechnen wir als mindestens notwendigen Stichprobenumfang für unsere Fragestellung n = 195. Aus (2) ist ersichtlich, n um so kleiner wird, je kleiner die mutmaßliche Varianz ist, oder je kleiner der Sicherheitsgrad (einem Sicherheitsgrad von 67% entspricht t = 1, von 99% t = 3); oder je größer der zulässige Fehler e genommen wird.

In unserem Beispiel wählen wir nun weitere 170 Sätze zufällig aus, da die 25 Sätze der Voruntersuchung in die für den Stichprobenumfang notwendigen 195 eingehen können, somit sich die Voruntersuchung als Teil der eigentlichen Untersuchung gestaltet.

Es ergibt sich bei der Stichprobe im Umfang von n = 195 Sätzen eine durchschnittlic Satzlänge von $\bar{x} = 19,83$. Ferner erhält man als Varianz der Stichprobenwerte um ihren Mittelwert nach

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} \qquad (3)$$

$s^2 = 146,27$. Die Verallgemeinerung dieses stichprobenartig gewonnenen Untersuchun ergebnisses, also der Rückschluß von einem Parameter der Erhebungsmasse auf einen

entsprechenden Parameter der Grundgesamtheit, erfolgt bei heterograder Fragestellung nach

$$\mu = \bar{x} \pm t \cdot \bar{x} \quad , \text{ wobei} \qquad (4)$$

$$\sigma^2_{\bar{x}} = \frac{s^2}{n} \cdot \frac{N-n}{N-1} \qquad (5)$$

Hierbei ist das s^2 in (5) jenes, welches aus (3) gewonnen wird.
Um zum Beispiel zurückzukehren: Wir erhalten für μ die durchschnittliche Satzlänge von TEMP, bei 95% Sicherheitsgrad:

$$\mu = 19,83 \pm 1,96 \cdot 0,5744 \quad \text{nach (5)}$$
$$18,71 \leq \mu \leq 20,95$$

Mit einer Sicherheit von 95% wird sich das wahre Ergebnis für die durchschnittliche Satzlänge von TEMP innerhalb der errechneten Grenzen befinden. Bei einer Totalerhebung hat sich für die durchschnittliche Satzlänge von TEMP $\mu = 19,59$ ergeben. Der bei dieser Stichprobe ermittelte Fehler $e = 1,12$ ist wesentlich geringer als der in der Aufgabenstellung zugelassene Fehler von 1,5. Dies liegt darin begründet, daß das errechnete $s^2 = 146,27$ weit unter dem bei der Berechnung des notwendigen Stichprobenumfang eingesetzten $s^{*2} = 200$ liegt, somit also der Stichprobenumfang größer ausfiel, als er hätte sein müssen, um das Ergebnis innerhalb einer Fehlerspanne von $\pm 1,5$ zu erhalten. Man hätte folglich mit einem geringerem Stichprobenumfang auskommen können. Der zugrunde gelegte Stichprobenumfang gibt also das Ergebnis innerhalb einer kleineren Fehlerspanne an.

2.2.2. Uneingeschränkte Zufallsauswahl homograder Natur

Prinzipiell ist das Vorgehen bei einer Stichprobe unter Verwendung einer Zufallsauswahl bei homograder Fragestellung dasselbe wie im eben beschriebenen Falle. Nur erhalten wir als Ergebnis der Voruntersuchung keinen Durchschnittswert $\bar{x}^*$,

sondern einen Anteilswert p^*, zum Beispiel für solche Finita, die im Konjunktiv stehen, und einen Anteilswert $q^* = (100 - p^*)$ für die Finita, die nicht im Konjunktiv stehen. Der Schätzwert für die mutmaßliche Varianz s^{*2} bemißt sich nach

$$s^{*2} = p^* \cdot q^* = p^* \cdot (100 - p^*) \qquad (6)$$

Mit diesem Wert für die mutmaßliche Varianz der Stichprobenanteilswerte kann ma nach Gleichung (2) den notwendigen Stichprobenumfang berechnen, wobei der zugelassene Fehler e das ± an Prozenten wiedergibt, welches man dem Endergebnis höchstens bei einer festgesetzten Sicherheit der Aussage beimessen möchte. Man kann sich aber auch eine Voruntersuchung ersparen, da sich im homograd en Fall eine höchstmögliche Varianz angeben läßt (der Wert für s^2 als Produkt zweier Anteilswerte nimmt dann den größtmöglichen Wert an, wenn $p = q = 50\%$).

Ist die Stichprobe durchgeführt, so erhält man einen Wert für die Varianz der Anteilswerte in der Stichprobe :

$$s^2 = p \cdot q = p \cdot (100 - p) \qquad (7)$$

Der Rückschluß von den Stichprobenanteilswerten auf Anteilswerte der Grundgesamtheit erfolgt analog zum heterograden Fall :

$$P = p \pm t \cdot \sigma_p \text{ , wobei} \qquad (8)$$

$$\sigma_p^2 = \frac{s^2}{n} \cdot \frac{N - n}{N - 1} \qquad (9)$$

2.2.3. Uneingeschränkte Zufallsauswahlen bei sehr großem oder unbekannt großem Umfang der Population

Fragen wir uns, wie sich die Ermittlung des notwendigen Stichprobenumfangs sowie der Rückschluß von Parametern der Stichprobe auf entsprechende der Grundgesamtheit gestaltet, wenn der Umfang der Grundgesamtheit im Vergleich zur angestrebten Stichprobe sehr große ist, oder wenn der Umfang der Grundgesamtheit als unbekannt groß angesehen werden muß. In solchen Fällen nimmt die Gliederung für die Berechnung des notwendigen Stichprobenumfangs folgende Form an:

$$n \leqq \frac{t^2 s^2}{e^2} \qquad (10)$$

Das für den Rückschluß notwendige $\sigma_{\bar{x}}$, beziehungsweise σ_p errechnet sich nach:

$$\sigma_{\bar{x}}^2, \text{ bzw. } \sigma_p^2 = \frac{s^2}{n}, \qquad (11)$$

wobei im heterograden Fall das s^2 sich nach Gleichung (3), im homograden Fall nach Gleichung (7) bestimmt.

2.2.4. Eingeschränkte Zufallsauswahl im Gegensatz zu uneingeschränkter Zufallsauswahl

Die bisher beschriebenen uneingeschränkten Zufallsauswahlen homograder und heterograder Natur bieten sich an, wenn es gilt, über eine in sich recht homogene Gesamtheit Aussagen zu machen, ohne auf eine Totalerhebung zurückgreifen zu müssen. Ist die Grundgesamtheit jedoch unhomogen, vielschichtig - dies kann bewirken, daß die Varianz der Stichprobenwerte um den Stichprobenmittelwert, das s^2, sich so groß gestaltet, daß der Rückschluß auf einen Parameter der Grundgesamtheit nur innerhalb eines nicht befriedigenden Fehlerintervalls möglich ist - , so bietet sich statt der uneingeschränkten Zufallsauswahl

ein Stichprobenverfahren mit eingeschränkter Zufallsauswahl an. Ein solches Verfahren ist die geschichtete Stichprobe. Sie hat gegenüber der uneingeschränkten Zufallsauswahl den Vorteil, daß sie bei gleichem Erhebungsumfang das Ergebnis für einen Wert der Grundgesamtheit innerhalb engerer Grenzen zeigt, beziehungsweise daß man bei gleich gehaltenen Grenzen mit einem geringerem Stichprobenumfang auskommt. Dies ist möglich, weil bei der geschichteten Stichprobe der Umfang von n Erhebungseinheiten nicht wie bei der uneingeschränkten Zufallsauswahl auf einmal zufällig gegriffen wird, sondern vielmehr als Summe einzelner Zufallsauswahlen. Im weiteren soll die geschichtete Stichprobe allgemein beschrieben und an einem Beispiel verdeutlicht werden.

2.2.5. Die geschichtete Stichprobe

Bei der uneingeschränkten Zufallsauswahl haben wir aus einer Grundgesamtheit von N Untersuchungselementen n Untersuchungselemente herausgegriffen. Bei einer geschichteten Stichprobe wird vor dem stichprobenartigen Erfassen der Untersuchungselemente die Population in K Schichten geteilt. Somit teilt sich die Anzahl der Untersuchungseinheiten der Grundgesamtheit N auf einzelne Schichten auf, wobei die j-te Schicht N_j Untersuchungselemente enthält. Es muß die Summe aller Schichtumfänge gleich sein dem Umfang der ungeschichteten Grundgesamtheit. Aus jeder Schicht der Grundgesamtheit wird nun eine uneingeschränkte Zufallsauswahl genommen, wodurch die Stichprobe ebenfalls in K Schichten geteilt wird (jeweils die j-te Schicht hat einen Umfang von n_j Untersuchungselementen). Die Summe aller Stichprobenumfänge für die einzelnen Schichten der Grundgesamtheit ist gleich dem Gesamtumfang der Stichprobe.
Es gilt:

	Grundgesamtheit	Stichprobe
Anzahl der Schichten	K	K
Anzahl der Untersuchungselemente in der j-ten Schicht	N_j	n_j
Gesamtzahl der Untersuchungseinheiten	$N = \sum_{1}^{K} N_j$	$n = \sum_{1}^{K} n_j$

Die Berechnung eines bestimmten Zahlenwertes für die Grundgesamtheit setzt voraus, daß man zunächst für jede einzelne der K Stichproben einen Mittelwert $\bar{x}_i$, beziehungsweise einen Anteilswert p_i nebst der dazugehörigen Varianz s_i^2 ermittelt. Der Mittelwert, beziehungsweise der Anteilswert der Stichprobe, ergibt sich dann nach

$$\bar{x} = \frac{\sum N_i \bar{x}_i}{N} \qquad (12)$$

$$p = \frac{\sum N_i p_i}{N} \qquad (13)$$

Der Rückschluß vom Mittelwert, vom Anteilswert der Stichprobe auf den entsprechenden Wert der Grundgesamtheit erfolgt nach Gleichung (4), beziehungsweise im homograden Fall nach Gleichung (8). Hierbei gilt bei der Anwendung einer geschichteten Stichprobe für $\sigma_{\bar{x}}^2$; beziehungsweise für σ_p^2 die folgende Gleichung:

$$\sigma_{\bar{x}}^2, \text{ bzw. } \sigma_p^2 = \frac{1}{N^2} \cdot \sum N_i^2 \cdot \frac{s_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} \qquad (14)$$

wobei das s_i^2 für die einzelnen Stichproben aus den Schichten nach Gleichung (3), im homograden Fall nach Gleichung (7) ermittelt wird.

Der gesuchte Parameter der Grundgesamtheit wird in umso engeren Grenzen getroffen, das heißt, das Vertrauensintervall wird um so geringer, je kleiner der Wert $\sigma_{\bar{x}}$, beziehungsweise σ_p wird; dies ergibt sich direkt aus den Gleichungen (4) und (8), die ja für den Rückschluß gelten. Für die geschichtete Stichprobe bedeutet dies, daß der Wert, den man aus Gleichung (14) erhält, sich möglichst klein gestalten soll. Aus Gleichung (14) ist ersichtlich, daß der aus ihr resultierende Wert um so kleiner wird, je kleiner sich die einzelnen s_i^2 gestalten.

Die s_i^2 geben die Varianz der Untersuchungsresultate einer Schicht um den zugehörigen Schichtmittelwert an. Wenn nun die Untersuchungsergebnisse einer Schicht möglichst wenig um ihren Mittelwert streuen, so ist die Varianz dieser Schicht recht gering. Teilt man folglich die Grundgesamtheit dergestalt in Schichten, daß innerhalb jeder Schicht eine möglichst geringe Varianz beobachtet werden kann, so erreicht man, daß aus Gleichung (14) ein möglichst kleiner Wert resultiert, was, wie aufgezeigt, dazu führt, daß die Aussage über die Grundgesamtheit innerhalb möglichst enger Fehlergrenzen erfolgen kann. Der Effekt der Schichtenbildung liegt also darin, mutmaßlich gleichgeartete Untersuchungselemente zu einer Schicht zusammenzufassen, wobei die Schichtung um so wirkungsvoller wird, je homogener man die einzelnen Schichten zu gestalten vermag.

Gerade ein solches Verfahren bietet sich als Hilfsmittel in der Sprachforschung an. Man kann bei verschiedenen Untersuchungen am Corpus des Instituts für deutsche Sprache immer wieder feststellen, daß innerhalb eines zum Corpus gehörigen Werkes relativ homogene Verhältnisse herrschen, die Werke unter sich betrachtet jedoch eine recht unhomgene Masse bilden. Es liegt nahe, das Corpus als Grundgesamtheit zu betrachten, welches in soviele Schichten geteilt werden kann, wie es Werke umfasst. Man wird bei gleichem Arbeitsaufwand, das heißt bei einem festgelegten Untersuchungsumfang n, wesentlich genauere Aussagen über das Corpus machen können, wenn man die n Untersuchungseinheiten nicht einfach zufällig aus dem Corpus greift, sondern die Gesamtstichprobe aufteilt auf einzelne Stichproben aus den einzelnen Werken (Schichten).

Es erhebt sich sofort die Frage, wie man bei einer geschichteten Stichprobe den insgesamt notwendigen Stichprobenumfang ermittelt, und wie man diesen notwendigen Umfang auf die einzelenen Stichproben aus den Schichten verteilt.
Um diese Frage zu beantworten, ferner um ein Beispiel der Anwendung einer geschichteten Stichprobe zu geben - und nicht zuletzt um auch zu zeigen, wieviel Untersuchungsarbeit man sich durch stichprobenartige Untersuchungen des Corpus ersparen kann - sei die folgende Problemstellung gegeben.

2.2.6. Die geschichtete Stichprobe mit prozentualer Aufteilung

Ermittelt werden soll die Gebrauchshäufigkeit des Konjunktivs. Der Untersuchung liegt ein Corpus von 11 Werken zugrunde. Bezugsgröße sei die Zahl der finiten Verben.

Um sich einen Überblick zu verschaffen, wieviele Finita man höchstens zu untersuchen hat, könnte man die Untersuchung auf einer uneingeschränkten Zufallsauswahl aufbauen. Wir haben es mit einer Fragestellung homograder Natur zu tun. Nach Gleichung (10) ließe sich der notwendige Stichprobenumfang ermitteln. Um sich eine Voruntersuchung zu ersparen, kann man in Gleichung (10) den maximalen Wert für s^2 einsetzen. Weiter sei angenommen, man wolle das Ergebnis bei 95%iger Sicherheit auf 1% genau ermitteln. Mit diesen Werten, eingesetzt in Gleichung (10), beläuft sich der notwendige Stichprobenumfang auf 10 000 Finita. Da aber sicherlich die Gebrauchshäufigkeit des Konjunktivs nicht bei 50% liegt, das heißt der Konjunktiv nicht bei jedem zweiten Finitum vorkommt, sondern wesentlich seltener, wird man schon bei geringerem Stichprobenumfang das Ergebnis innerhalb der gewünschten Grenzen bei gesetztem Sicherheitsgrad bekommen.

Will man eine geschichtete Stichprobe durchführen, so muß man neben dem Umfang der Grundgesamtheit auch noch die Umfänge der einzelnen Schichten kennen, in welche die Grundgesamtheit aufgeteilt wurde. In unserem Beispiel muß bekannt sein, wieviele Finita das gesamte Corpus und jedes einzelne Werk enthält, wenn man davon ausgeht, daß jedes Werk eine Schicht bilden soll. Da eine exakte Auszählung zu arbeitsreich wäre, genügt es, die Finitazahlen der Werke mittels uneingeschränkten Zufallsauswahlen abzuschätzen, wobei die sich ergebende Obergrenze der Finitazahl eines Werkes in die weiteren Berechnungen eingehen soll. Das Ergebnis lautet, die i-te Schicht beinhaltet N_i Untersuchungseinheiten (Finita). Durch die Summe aller N_i ist der Umfang der Grundgesamtheit gegeben. Man weiß nun auch, wie sich N auf die einzelnen N_i verteilt.

Führt man eine geschichtete Stichprobe durch, so erhebt sich die Frage, wie der Stichprobenumfang n sich auf die einzelnen n_i verteilen soll. Eine gängige Methode der Aufteilung ist die prozentuale : Gemäß dem Anteil einer N_i umfangreichen Schicht

an der Grundgesamtheit bemißt sich der Anteil der n_i umfangreichen Schichtstichprobe an dem Gesamtumfang n der Stichprobe. Die einzelnen n_i sind also proportional zu den N_i zu nehmen. Folglich muß bei einer solchen prozentualen Aufteilung gelten:

$$n_i = c \cdot N_i \text{ , wobei } c = \text{const.} = \frac{n}{N} \qquad (15)$$

Das n berechnet sich bei einer geschichteten Stichprobe mit prozentualer Aufteilung nach Gleichung (16), wobei ein maximal zulässiger Fehler und der Sicherheitsgrad zu setzen sind:

$$n = \frac{t^2 \cdot N \cdot \sum N_i s_i^2}{N^2 \cdot e^2 + t^2 \sum N_i s_i^2} \qquad (16)$$

Werke i =	1	2	3	4	5	6	7	8	9	10	11
N_i	1200	10500	8000	5000	6700	5000	10200	1200	2700	9700	3000
n_i	120	1050	800	500	670	500	1020	120	270	970	300
p_i	12,25	10,40	5,65	5,25	4,61	2,00	3,70	10,00	5,80	7,00	35,30

In obiger Tabelle sind die Untersuchungsergebnisse eingetragen, ferner ist ersichtlich daß für die prozentuale Aufteilung c = o,1 angenommen wurde.
Freilich läßt sich c auch rechnerisch ermitteln nach Gleichungen (15) und (16), wenn man den Sicherheitsgrad und den Fehler setzt. Sollte in unserem Beispiel die Gebrauchshäufigkeit des Konjunktivs auf 1 % genau ermittelt werden, wobei der höchst zulässige Irrtum bei der Ergebnisfindung 5 % betrage, so errechnet sich der notwendige Stichprobenumfang, wenn man für die einzelnen s_i^2 den ungünstigsten

Wert 0,25 annimmt, nach Gleichung (15) : n = 8 633 (man könnte auch statt des ungünstigsten Wertes für s_i^2 durch eine Voruntersuchung einen Schätzwert s_i^{*2} gewinnen und in die Rechnung für den notwendigen Stichprobenumfang eingehen lassen).

Wird n nach Gleichung (16) errechnet, so ergibt sich c als Quotient von n und N. Wird c wie in unserem Beispiel gesetzt, so ergibt sich n als Produkt von c und N. Ist c gesetzt, so kann n ebenfalls als gesetzt gelten. Mit diesem gesetzten n kann man in Gleichung (16) dann bei gegebenem Sicherheitsgrad den größtmöglichen Fehler bestimmen, der bei diesem gesetzten n auftreten kann, indem man Gleichung (16) nach e zur Auflösung bringt. Es liegt dann im eigenem Ermessen, ob man den maximal möglichen Fehler für den Parameter der Grundgesamtheit als noch gängig erachtet, oder ob man ihn kleiner gestalten möchte, indem man den Stichprobenumfang erhöht.

Dann beginnt die Untersuchung. Es wird der c-te Teil jeder Schicht vom Umfang N_i Untersuchungseinheiten uneingeschränkt zufällig erhoben. In unserem Beispiel werden also n_i Finita zufällig aus den einzelnen Werken gegriffen, damit festgestellt werden kann, wieviel Prozent der Finita einer Schicht im Konjunktiv stehen. Ergebnisse siehe Tabelle.

Nach Gleichung (13) berechnet sich ein p = 7,523 %. Der Rückschluß auf den Wert der Grundgesamtheit erfolgt nach Gleichung (8), wobei für $\sigma_{\bar{x}}$ bzw. σ_p bei einer geschichteten Stichprobe mit proportionaler Aufteilung gilt :

$$\sigma_{\bar{x}}^2 \text{ bzw. } \sigma_p^2 = \frac{N-n}{N^2 \cdot n} \cdot \sum N_i s_i^2 \qquad (17)$$

In unserem Beispiel wird σ_p nach Gleichung (17) bestimmt : σ_p = 0,30436 %. Somit können wir als Aussage für den Wert der Grundgesamtheit wiedergeben :

Sicherheitsgrad 95% : 6,915% $\leq$ P $\leq$ 8,131%

Was bedeutet nun dieses Ergebnis ? Bei einer Untersuchung von etwa 63 200 Finita aus 11 totalerhobenen Werken gelangt man natürlich zu einem exakten

Wert für die Gebrauchshäufigkeit des Konjunktivs : P = 7,523%[5]. Versucht man diesen Wert stichprobenartig zu bestimmen, so erhält man ihn schon bei einem Stichprobenumfang von nur 10% des Umfangs der Population bei einem Irrtumsrisiko von 5% auf ± 0,6% genau. Erscheint das Fehlerrisiko von 5% als zu groß, so kann man das Vertrauensintervall weiter fassen : Mit einem Sicherheitsgrad von 99,7% liegt der wahre Wert für P innerhalb einer Spanne von ± 0,9%. Wesentlich engere Grenzen erhält man natürlich, wenn der Stichprobenumfang vergrößert wird[6].

Warum es sinnvoll sein kann, Untersuchungen mit Hilfe von Stichprobenverfahren durchzuführen, ist nun deutlich geworden. 90% der Arbeitsaufwendungen für das Ermitteln von Daten und deren Weiterverarbeitung kann man einsparen, bzw. zu anderen Zwecken verwenden, wenn man sich damit zufrieden gibt, anstatt eines exakten Wertes, einen Wert zu erhalten, der innerhalb eines genau bestimmbaren maximalen Fehlerbereiches liegt[6]. Ist grundsätzlich die Bereitschaft vorhanden, sich mit einem Ergebnis innerhalb eines Vertrauensbereiches zu begnügen, so läßt sich berechnen, wieviel Prozent der Grundgesamtheit man erheben muß. Diesem Sachverhalt liegt als fundamentaler Zusammenhang Gleichung (10) zugrunde : $n = t^2 \cdot s^2/e^2$.

Setzt sich der Untersuchende eine obere Grenze für den Stichprobenumfang, so kann er abschätzen, innerhalb welcher Fehlergrenzen bei gegebenen genügend hohem Sicherheitsgrad das Endergebnis liegen wird. Setzt sich der Untersuchende aber eine Fehlergrenze, die er als äußerste zu dulden bereit ist, so kann er den hierfür notwendigen Umfang der Stichprobe abschätzen. Wie groß also letztlich der Stichprobenumfang 'n' gewählt wird, beruht auf einem Kompromiß zwischen dem zur Verfügung stehenden Arbeitsaufwand und der gewünschten Genauigkeit. Wenn man nun, um das Konjunktiv-Beispiel abzuschließen, tatsächlich die Möglichkeit hat, 63 200 Finita zu untersuchen, so erscheint es sinnvoller, statt aus 11 Werken Totalerhebungen durchzuführen, das Corpus wesentlich zu erweitern und wiederum eine stichprobenartige Erhebung zu machen, wobei man zwar wieder nur Werte erhält, die innerhalb eines beliebig bestimmbaren Fehlerbereiches liegen; aber diese Werte sind in ihrer Aussagekraft für die geschriebene Sprache von ungleich größerer Bedeutung[7].

3. Allgemeines zur Anwendung der geschichteten Stichprobe bei Corpusuntersuchungen

Das Corpus stellt sich uns als eine nach Texttypen gegliederte Ansammlung von Texten dar. Es ist eine Teilerhebung nach Gesichtspunkten der Repräsentativität aus einer bestimmt definierten übergeordneten Gesamtheit zum Beispiel der geschriebenen Sprache der Gegenwart. Jede der nach numerischen oder nichtnumerischen Gesichtspunkten gebildeten Texttype kann als Schicht verstanden werden[8]. Die Schichtenbildung, als notwendige Voraussetzung zur Anwendung einer geschichteten Stichprobe erfolgt also analog zur Typenbildung.

So gilt allgemein für Corpus und Stichprobe:

	Corpus	Stichprobe
Anzahl der Schichten	M	M
Anzahl der Texte in der j-ten Schicht	I	I
Gesamtzahl der Untersuchungseinheiten des i-ten Textes in der j-ten Schicht	$\sum N_{ji}$	$\sum n_{ji}$
Gesamtzahl der Untersuchungseinheiten der j-ten Schicht	$N_j = \sum_{i=1}^{I} N_{ji}$	$n_j = \sum_{i=1}^{I} n_{ji}$
Gesamtzahl der Untersuchungseinheiten des Corpus	$N = \sum_{1}^{M} N_j = \sum_{1,1}^{I,M} N_{ji}$	$n = \sum_{1}^{M} n_j = \sum_{1,1}^{I,M} n_{ji}$

Der notwendige Umfang der Stichprobe aus dem Corpus läßt sich auf Grund der Gleichung (16) rechnerisch bestimmen. Um diese Gleichung anwenden zu können, bedarf es zur Bestimmung der einzelnen s_i^2 einer Voruntersuchung, welche man sich bei Fragestellungen homograder Natur ersparen kann, da dann, wie beschrieben, der maximale Wert jeweils für die einzelnen s_i^2 angenommen werden kann. Andererseits ist die Voruntersuchung nicht nur für die Berechnung des notwendigen Stichprobenumfangs von Nutzen, wenn diese im Rahmen des vorher fixierten Auswahlverfahrens durchgeführt wird, denn dann kann sie als Teil der eigentlichen Stichprobenerhebung angesehen und verwendet werden. Wie eben beschrieben, kann man sich den Stichprobenumfang setzen, etwa als diejenige Anzahl Untersuchungseinheiten, die man auf Grund bestimmt großer möglicher Arbeitsaufwendung zu untersuchen in der Lage ist. Dann läßt sich, wie oben beschrieben, nach Gleichung (16) die Fehlergrenze ermitteln, innerhalb dieser man das Ergebnis bei gesetztem Sicherheitsgrad erhalten wird.
Ist n bekannt, so ist gleichzeitig der Umrechnungsfaktor c für die proportionale Aufteilung des Stichprobenumfangs gegeben als Quotient von N und n. In die Stichprobe geht der c-te Teil jedes Corpus-Schichtumfanges ein.

Es sei nochmals betont, daß das Auswahlverfahren so bestimmt werden muß, daß nur der Zufall bestimmt, welche die durch j und i indizierte Untersuchungseinheit eines bestimmten Textes in die Schichtstichprobe vom Umfang n_i gelangt.

Die Ergebnisse der stichprobenartigen Untersuchung des Corpus können nach den angeführten Gleichungen verallgemeinert werden. Man erhält somit einen Vertrauensbereich, innerhalb dessen der wahre Wert, also der Wert, der durch eine Totalerhebung des Corpus zu gewinnen wäre, mit festgelegter Sicherheit liegen wird.

Ob das Ergebnis auch Gültigkeit für die dem Corpus übergeordnete Gesamtheit hat ob es also als repräsentativ z.B. für die geschriebene Sprache der Gegenwart erachtet werden kann, hängt ab vom Grad der Repräsentativität den man dem Corpus beimißt[9].

Anmerkungen

1) Eine Darstellung von Theorie und Technik der Stichprobenverfahren findet sich in Kellerer, 1963.
Die Beispiele entstammen Untersuchungsarbeiten des Instituts für deutsche Sprache.

2) Königerova, 1966, behandelt ebenfalls die Anwendung der Stichprobentheorie auf die Sprachbearbeitung, sie beschränkt ihre Ausführungen auf die Anwendung uneingeschränkter Zufallsauswahlen.

3) Kellerer, 1966, S. 14.

4) TEMP kennzeichnet einen Corpustext: Bergengruen, Das Tempelchen.

5) Dieser Wert beruht auf Untersuchungen von Jäger, der im Institut für deutsche Sprache an einer Dokumentation des Konjunktives arbeitet.

6) Freilich gilt es zu bedenken, daß bei einer Stichprobenerhebung sehr seltene Erscheinungen gar nicht wahrgenommen werden, weil sie zufällig nicht in die Stichprobe gelangt sind. Zum Beispiel kann Jäger besondere Formen des Konjunktives bei Untersuchung von über 60 000 Finita nur drei oder viermal belegen. So ergibt sich die Frage, ob das Entdecken äußerst seltener Formen eine wesentlich umfangreichere Erhebung rechtfertigt.

7) Mit diesen hier beschriebenen Methoden lassen sich natürlich nur textlängenunabhängige Zahlen und Zahlenverhältnisse stichprobenartig gewinnen. Textlängenabhängige Größen, wie zum Beispiel die Wiederholungsrate der Types oder die Silbenzahl der Types, werden mittels Regressionsmethoden geschätzt; vgl. Müller, wo die Bestimmung von textlängenabhängigen Textparametern eines Textes durch Untersuchung eines Teiltextes gezeigt ist.

8) Zur Einteilung von Texten in Texttypen auf Grund numerischer Gesichtspunkte, vgl. Müller.

9) Über Möglichkeiten zur Quantierung dieses Problems vgl. Müller.

Anmerkungen

1) Eine Darstellung von Theorie und Technik der Stichprobenverfahren findet sich in Kellerer, 1963.
Die Beispiele entstammen Untersuchungsarbeiten des Instituts für deutsche Sprache.

2) Königerova, 1966, behandelt ebenfalls die Anwendung der Stichprobentheorie auf die Sprachbearbeitung, sie beschränkt ihre Ausführungen auf die Anwendung uneingeschränkter Zufallsauswahlen.

3) Kellerer, 1966, S. 14.

4) TEMP kennzeichnet einen Corpustext: Bergengruen, Das Tempelchen.

5) Dieser Wert beruht auf Untersuchungen von Jäger, der im Institut für deutsche Sprache an einer Dokumentation des Konjunktives arbeitet.

6) Freilich gilt es zu bedenken, daß bei einer Stichprobenerhebung sehr seltene Erscheinungen gar nicht wahrgenommen werden, weil sie zufällig nicht in die Stichprobe gelangt sind. Zum Beispiel kann Jäger besondere Formen des Konjunktives bei Untersuchung von über 60 000 Finita nur drei oder viermal belegen. So ergibt sich die Frage, ob das Entdecken äußerst seltener Formen eine wesentlich umfangreichere Erhebung rechtfertigt.

7) Mit diesen hier beschriebenen Methoden lassen sich natürlich nur textlängenunabhängige Zahlen und Zahlenverhältnisse stichprobenartig gewinnen. Textlängenabhängige Größen, wie zum Beispiel die Wiederholungsrate der Types oder die Silbenzahl der Types, werden mittels Regressionsmethoden geschätzt; vgl. Müller, wo die Bestimmung von textlängenabhängigen Textparametern eines Textes durch Untersuchung eines Teiltextes gezeigt ist.

8) Zur Einteilung von Texten in Texttypen auf Grund numerischer Gesichtspunkte, vgl. Müller.

9) Über Möglichkeiten zur Quantierung dieses Problems vgl. Müller.

Das Mannheimer Corpus

von Ulrich Engel

Da sich die in diesem Forschungsbericht enthaltenen Beiträge ausdrücklich oder implizit mit den vom Institut für deutsche Sprache zusammengestellten und für weitere elektronische Bearbeitung auf Magnetband gespeicherten Texten befassen, da scheint es angebracht, näher auf Geschichte und Motivation des "Mannheimer Corpus" einzugehen.

Dieses Corpus (s. Seite 76 - 78) beruht auf Vorschlägen der wissenschaftlichen Mitarbeiter des Instituts, die in mehreren Sitzungen vom Wissenschaftlichen Rat und den dafür zuständigen Kommissionen (Dokumentation, Grunddeutsch) erörtert und ergänzt wurden.

Es wurde eine Auswahl angestrebt, die als repräsentativ angesehen werden konnte für die deutsche Gegenwartssprache (seit 1945) in ihrer geschriebenen Form, wobei von regionalen und sozialen Dialekten und anderen Sonderformen mit beschränktem Geltungsbereich abgesehen wurde.

Damit war von vornherein klar, daß das Corpus nicht auf die sogenannte schöne Literatur beschränkt werden durfte. Im ganzen schälten sich drei große Bereiche heraus: Neben der schönen Literatur einschließlich ihrer Trivialformen vor allem die Fachliteratur, die sowohl in Einzelbänden als auch in mehreren Zeitschriften erfaßt wurde; schließlich die Sprache der politischen Nachrichten, wie sie sich im wesentlichen auf Seite 1 der Tageszeitungen finden. Die "Bild Zeitung" freilich ist als ein Phänomen sui generis zu betrachten, zweifellos enthält sie nicht einfach einen spezifischen Aspekt von Zeitungssprache.

Eine wichtige Rolle spielte die Frage, was tatsächlich gelesen wird. Aus dieser Fragestellung unter anderem[1)] erklärt sich die Einbeziehung der Trivialliteratur mit 3 Werken, die starke Bevorzugung populärwissenschaftlicher Werke gegenüber der Fachprosa im engeren Sinne, schließlich die Aufnahme eines halben Jahrganges der Bild Zeitung.

Deutsche Textbibliothek

Maschinell gespeicherte Texte des Instituts für deutsche Sprache

Bergengruen, Werner, Das Tempelchen, Erzählung, Arche, Zürich, Nymphenburger Verlagshandlung, München (C. 1950 Peter Schifferli, Verlags AG Die Arche, Zürich).

Böll, Heinrich, Ansichten eines Clowns, Roman, Kiepenheuer und Witsch, Köln - Berlin, 113.- 157. Tausend August 1964 (1. - 28. Tausend Mai 1963).

Frisch, Max, Homo Faber, Ein Bericht, Bibliothek Suhrkamp, Bd. 87, 161.-180. Tausend 1966 (C. 1957 Suhrkamp, Frankfurt/Main).

Grass, Günter, Die Blechtrommel, Roman, Firscher, Frankfurt/Main und Hamburg, 323.- 372. Tausend Mai 1964 (1.-50. Tausend September 1962) (C. 5. und 6. Auflage August 1960 Luchterhand, Darmstadt - Berlin - Neuwied).

Johnson, Uwe, Das dritte Buch über Achim, Roman, Suhrkamp, Frankfurt/Main 1961, 16.-20. Tausend (C.1961).

Mann, Thomas, Die Betrogene, Erzählung, S. Fischer, Frankfurt/Main, 16.- 20. Tausend 1954 (C. 1953).

Strittmatter, Erwin, Ole Bienkopp, Roman, Sigbert Mohn, Gütersloh (C. 1963 Aufbau-Verlag, Berlin W8).

Jung, Else, Die Magd vom Zellerhof, Kelter Heimat-Roman Bd. 41, Martin Kelter, Hamburg-Wandsbeck (o.J.).

Pinkwart, Heinz, Mord ist schlecht für hohen Blutdruck, Kriminalroman, Goldmann, München (C. 1963).

Stauffen, Pia, Solange dein Herz schlägt, Juwelen-Roman Nr. 748, Pabel, Rastatt (Baden) (o.J.).

Bamm, Peter, Ex Ovo, Essays über die Medizin, Deutsche Verlags-Anstalt, Stuttgart, 63.- 65. Tausend 1963 (C. 1956).

Bollnow, Otto Friedrich, Maß und Vermessenheit des Menschen, Philosophische Aufsätze, Neue Folge, Vandenhoeck & Ruprecht, Göttingen und Zürich (C. 1962).

Gail, Otto Willi und Petri, W., Weltraumfahrt, Physik - Technik - Biologie, 2., völlig neubearbeitete Auflage des Werkes Physik der Weltraumfahrt , 1947, Hanns Reich, München (C. 1958).

Grzimek, Bernhard, Serengeti darf nicht sterben, Ullstein, Berlin, 131.- 141. Tausend April 1963 (1.-30. Tausend September 1959).

Heimpel, Hermann, Kapitulation vor der Geschichte?, 3., vermehrte Auflage, Vandenhoeck & Ruprecht, Göttingen, 13.-18. Tausend 1969 (C. 1956).

Heisenberg, Werner, Das Naturbild der heutigen Physik, rde 8, Rowohlt, Hamburg, 124. - 128. Tausend September 1966 (1.- 40. Tausend Dezember 1955), bis S. 46.

Jaspers, Karl, Die Atombombe und die Zukunft des Menschen, Piper, München, 37.- 44. Tausend 1962 (C. 1958).

Jungk, Robert, Die Zukunft hat schon begonnen, Amerikas Allmacht und Ohnmacht, neue erweiterte Ausgabe, Scherz, Bern - München - Wien (C. 1952).

Pörtner, Rudolf, Die Erben Roms, Städte und Stätten des deutschen Früh-Mittelalters, Econ, Düsseldorf - Wien, 41.- 70. Tausend 1965 (1.- 40. Tausend 1964).

Staiger, Emil, Grundbegriffe der Poetik, Atlantis, Zürich - Freiburg/Breisgau 1966, 7. Auflage (C. 1946).

Ullrich, Fritz, Wehr Dich Bürger! Aktuelle Rechtsschutzfibel, Gieseking, Bielefeld (C. 1960).

Heuß, Theodor, Erinnerungen, 1905-1933, Wunderlich, Tübingen, 5. Auflage, 71.- 85. Tausend Mai 1964 (1. Auflage September 1963).

"Frankfurter Allgemeine Zeitung", D-Ausgabe, 19.1.1966 - 17.2.1966, jeweils die erste Seite (ohne Leitartikel).

"Die Welt", Ausgabe D***, 1.12.1965 - 18.2.1966, jeweils die erste Seite.

"Bild der Wissenschaft", hg. von Prof. Dr. Heinz Haber in der Deutschen Verlags-Anstalt, Stuttgart, Heft 1, 2 und 3/1967 (jedes Heft enthält 5 Artikel).

"Studium Generale", Schriftleitung G.G. Grau, Springer, Berlin - Heidelberg - New York, Heft 12/1966 (das Heft enthält 6 Artikel).

"Urania", hg. vom Präsidium der URANIA (Gesellschaft zur Verbreitung wissenschaftlicher Kenntnisse) und dem Deutschen Kulturbund, Urania, Leipzig - Jena - Berlin, Heft 11/1966 und 1/1967 (jedes Heft enthält 14 Artikel).

"Bild Zeitung" 7 Monate (Januar-Juli 1967), im Turnus 1.Tag/1.Seite, 2. Tag/2.Seite ... 7. Tag/1.Seite usw.

Auch der regionale Gesichtspunkt wurde berücksichtigt. Die Autoren vor allem der Bereiche "Schöne Literatur" und "Trivialliteratur" stammen aus verschiedenen Teilen des deutschen Sprachraums, mit Max Frisch ist auch die deutschsprachige Schweiz vertreten, die Aufnahme eines österreichischen Autors ist vorgesehen. Da die Möglichkeit, daß sozial und politisch verschiedene Systeme auch zu sprachlicher Differenzierung führen, nicht von der Hand zu weisen ist[2)], wurde mit dem Ole Bienkopp auch das Werk eines anerkannten Schriftstellers aus der DDR aufgenommen.

Schließlich war auf die Erfassung möglichst verschiedener Stilarten zu achten. Nachdem geschriebene Gegenwartssprache definiert worden war als Sprache der Gesamtheit der nach 1945 entstandenen Werke, mußte den hier offenkundig vorhandenen Stilunterschieden Rechnung getragen werden. Es wäre unzulässige Simplifizierung, stilistische Spezifika einfach vom Alter des Autors abhängig zu machen oder gar schlicht auf den Unterschied der Generationen zurückzuführen. Es kann aber nicht übersehen werden, daß eine Reihe vorwiegend älterer Autoren einen weitgehend an der deutschen Klassik orientierten Stil pflegt, während die Jüngeren bewußt mit der Schultradition brechen - faktisch oft viel weniger, als offenbar beabsichtigt ist - und neue, freiere, elastischere, der Alltagssprache näherstehende Stilformen sich heranzubilden beginnen. Auf dem Hintergrund solcher Beobachtungen mag es verständlich werden, daß (neben Frisch, Grass, Johnson) Thomas Mann und Werner Bergengruen mit Spätwerken aufgenommen werden.

Natürlich waren auch Beschränkungen notwendig. Von der maschinellen Erfassung wie von der linguistischen Auswertung her waren äußere Grenzen gesetzt: der Grad der Zuverlässigkeit der Ergebnisse muß in einem vernünftigen Verhältnis zu der aufgewendeten Arbeit stehen. Ein Corpus im Umfang von insgesamt etwa 1,6 Millionen Wörtern - rund das Dreißigfache von Max Frischs "Homo Faber" - konnte einerseits als ausreichend gelten, andererseits erlaubt es noch eine einigermaßen vollständige Auswertung.

Es war schon die Rede davon, daß Schwerpunktbildung erforderlich war. Dies gilt vor allem für den Bereich der wissenschaftlichen Literatur. Hier wurde, aus schon dargelegten Gründen, eng Fachbezogenes ausgeschieden, weil hier - zumeist

beim Wortschatz - Sonderbildungen zu erwarten waren, die für das Gesamtbild der deutschen Gegenwartssprache von zweitrangiger Bedeutung sind. Außerdem wurde aber auch, und zwar mit ausdrücklicher Zustimmung der Vertreter des Goethe-Instituts, für dessen Arbeit dieser Zweig besonders bedeutsam ist, die Gesamtmenge der Lehr- und Fachbücher weggelassen. Selbst durch die Unterrichtsliteratur sehr niederen Niveaus käme nämlich schon ein umfangreicher Wortschatz sehr spezieller Art in das Corpus, während die generellen lexikalischen und syntaktischen Merkmale der Fach- und Wissenschaftssprache ebensogut in den (vorwiegend) populärwissenschaftlichen Werken und den berücksichtigten Zeitschriften vertreten sind.

Daß von den beiden Zeitungen "Die Welt" und "Frankfurter Allgemeine Zeitung" - die "Bild Zeitung" ist in mehrfacher Hinsicht als Sonderfall zu betrachten - nur die erste Seite aufgenommen wurde, hat wohldurchdachte Gründe. Es ging ja von Anfang an nicht um "Zeitungssprache", ein ohnehin höchst heterogenes Gebilde, dessen Legitimation als Gegenstand linguistischer Forschung noch erbracht werden müßte, sondern es ging um die Sprache der politischen Nachrichten, dessen Untersuchung Peter von Polenz dezidiert gefordert hatte[3]. Diese Sprache findet sich zwar auch an anderen Stellen normaler Tageszeitungen, aber es ist legitim, die erste Seite der Zeitung als stellvertretend für den sprachlichen Gesamtbereich herauszugreifen.

Es stellt sich auch die Frage, in welchem Umfang die einzelnen Werke zu erfassen seien. Statistiker empfehlen in solchen Fällen immer möglichst viele kleinere Teiltexte; ". . . daß statistisch gesehen 10 Prozent aus 100 Titeln immer besser sind als 100 Prozent aus 10 Titeln", betont auch Manfred Hellmann in Abschnitt 2.5. seines Berichts. Diese Forderungen bestehen grundsätzlich zu Recht, wenn es in erster Linie darum geht, in dem Auswahlcorpus eine größere Grundgesamtheit möglichst "maßstabsgetreu" zu repräsentieren. Wenn wir uns doch zur Aufnahme der ungekürzten Texte entschlossen haben, so hauptsächlich auf Grund der Erwägung daß unser Corpus für möglichst viele beliebige Fragestellungen geeignet sein sollte. Dazu gehören aber auch mögliche kompositorisch bedingte Erscheinungen, die nur im weitesten Kontextzusammenhang untersucht werden können und deshalb in Auswahltexten verlorengehen könnten.

Die gesamte Diskussion läuft letztlich auf die Frage hinaus, was wir unter der Repräsentativität eines Corpus verstehen. Daraus wiederum ergeben sich zwei Teilfragen :

1) Wofür soll das Corpus repräsentativ sein?
2) Im Hinblick worauf soll es repräsentativ sein?

1) Wir betrachten als Grundgesamtheit, die vom Corpus abgebildet werden soll, die gemeindeutsche (interregionale, intersoziale, überfachliche) Gegenwartsprosa. Bei Anlegung streng statistischer Maßstäbe müßten möglichst viele Merkmale dieser Grundgesamtheit im Corpus vertreten sein, und zwar jeweils entsprechend der Häufigkeit ihres Vorkommens.

Es ist aber bisher noch kein Verfahren entwickelt worden, das die Gewinnung eines solcherart repräsentativen Corpus ermöglicht, ohne daß die Grundgesamtheit en détail bekannt wäre (wahrscheinlich wird ein solches Verfahren auch nie bis zur Funktionsreife gedeihen). Wie die Fülle des deutschen Prosaschrifttums unserer Zeit systematisch erfaßt werden könnte (man denke nur an die Mengen täglich erscheinender Zeitungen, auf die Hellmann hinweist), ist ein vorderhand gänzlich ungelöstes Problem. Das hängt weitgehend zusammen mit der Tatsache, daß eine praktikable Typik des deutschen Schrifttums nicht existiert. Der Kernsche Entwurf bietet ein wertvolles, einleuchtendes Instrumentarium für Textbeschreibungen und damit auch für eine Typik der deutschen Gegenwartssprache. Kriterien für die Typenbildung bietet er nicht, er will das auch gar nicht : dies zeigt der dritte Teil seiner Studie, in der präexistente "Typen" durch Kombination mehr oder weniger zahlreicher Merkmale charakterisiert werden. Weitgehend ist man sich über bestehende Typen einig (die literarischen Gattungen und anderes werden meist unreflektiert mit verwendet), ohne ihre Konstituierung begründen zu können : eine allgemeine, verbindliche Typik muß noch geschaffen werden.

Hätten wir diese Typik, hätten wir mithin ein allumfassendes "Schubkastensystem", in das jeder Text auf Grund exakt feststellbarer Merkmale eingeordnet werden könnte (wobei simultane Einordnung in verschiedene "Schubkästen" inbegriffen wäre), so könnten wir tatsächlich ein quantitativ getreues Abbild der Grundgesamtheit liefern : ein übersichtliches Corpus, in dem alle "Typen" des Gegenwartsschrifttums

entsprechend der Häufigkeit ihres Vorkommens vertreten wären.

Es fragt sich, was damit gewonnen wäre. Die wöchentlichen Bestsellerlisten des "Spiegel" verzichten mit gutem Grund auf die Einbeziehung der absoluten Bestseller, der Lehrbücher jeglicher Art und Provenienz. Das rein quantitative verkleinerte Abbild einer strikt typisierten Gegenwartsliteratur enthielte einen alles andere überwiegenden Teil an didaktischer Literatur, der in seiner Monotonie nur wenige Aufschlüsse verspräche. Schon deshalb scheint uns ein Ausgehen von den Auflagenziffern nicht diskutabel. Das deutsche Gegenwartsschrifttum kann durch bloße Aufreihung alles in deutscher Sprache Gedruckten nicht adäquat erfaßt werden. Neben dem Ausstoß der Druckereien muß mindestens auch der Grad der Wertschätzung beim lesenden Publikum, eingeschlossen die mehr oder weniger etablierten, von Kritikern oder von der Schule verbreiteten Wertnormen, berücksichtigt werden. Zahlreiche weitere Gesichtspunkte kommen hinzu. Es scheint mir weit eher vertretbar, von einer solchen auf Grund einer Vielzahl relevanter Kriterien "gewichteten", wenngleich quantitativ nicht so exakt abgegrenzten Gesamtmenge auszugehen, als von einer bloßen Summation des irgendwo und irgendwi Gedruckten. Für diese mehrfach gewichtete Gesamtmenge allerdings soll unser Corpus repräsentativ sein insofern, als es seine wesentlichen Merkmale ebenfalls enthält. Das ist, wie die bisherigen grammatischen Untersuchungen erwiesen haben (es wurden zum Vergleich und zur Ergänzung weitere Werke beigezogen) in hohem Maße der Fall. Freilich fehlen viele sprachlichen Sonderausprägungen. Man muß aber bedenken, daß dieses Corpus in erster Linie für die Untersuchung grammatisch syntaktischer Erscheinungen erstellt wurde, die sich zum Teil durch sehr verschiede artige Texte hindurch nur unwesentlich ändern. Ein speziell für Wortschatzuntersuchungen zusammengestelltes Corpus[4] müßte naturgemäß ganz anders aufgebaut werden. Dies führt uns zu der Frage :

2) Im Hinblick worauf soll das Corpus "repräsentativ" sein? Die Antwort auf diese Frage hängt eben nicht bloß von der definierten Gesamtmenge ab, deren Merkmale sich im Corpus wiederfinden sollen, sondern in noch höherem Maße von den zu untersuchenden Erscheinungen. Darüber liegen im Institut umfangreiche Erfahrungen vor. Während sich etwa das Corpus für Satzstrukturen, Wortstellung und bestimmte

Tempora als viel zu umfangreich erwies, so daß jeweils nur mit Teilmengen gearbeitet wurde, mußten für eine relativ seltene Erscheinung wie das Passiv in erheblichem Umfang weitere Texte beigezogen werden. Strenggenommen müßte für jede spezielle Fragestellung ein eigenes Corpus zusammengestellt werden. Wie ein solches Corpus jeweils beschaffen sein muß, ist oft nicht im voraus auszumachen, sondern ergibt sich vielfach erst als Teilergebnis der Untersuchung. Insofern können die Mannheimer Arbeiten zur deutschen Grammatik Hinweise für spätere weiterführende Untersuchungen ergeben.

Wo es dann um die Gewinnung von Teilmengen aus dem einmal festgelegten Corpus geht, soll freilich möglichst exakt verfahren werden. Das von Werner Müller dargelegte Verfahren ist für solche Zwecke nützlich. Billmeiers Vorschlag (im Forschungsbericht Nr. 2) ist zunächst für Wortschatzuntersuchungen konzipiert; die Anwendung auf grammatische Fragestellungen dürfte erheblich komplizierter sein.

Wir fassen zusammen : Die Erstellung eines Corpus, das die gesamte deutsche Gegenwartsprosa exakt abbildet, ist heute noch auf lange Sicht unmöglich, und sie wäre auch aus verschiedenen Gründen kaum vertretbar. Da sich die jeweilige Fragestellung als wichtige Konstante bei der Corpuskonstitution erwiesen hat, kann es ein allgemein gültiges Corpus auch gar nicht geben, man müßte es denn in allen Teilen so umfangreich anlegen, daß die Auswertung alle Grenzen der Wirtschaftlichkeit sprengen würde. Das Corpus des Instituts für deutsche Sprache war immer nur als Rahmenkonzeption zu verstehen; es kann bei Bedarf ebenso erweitert wie nur partiell für die Auswertung herangezogen werden. Würde man heute noch einmal beginnen, so würde man gewiß manches ändern. Vor vier Jahren fehlten viele der heute vorliegenden Erfahrungen. Im ganzen freilich würde sich wahrscheinlich ein nicht allzu stark abweichendes Bild ergeben. Wir halten dieses Corpus für repräsentativ für die deutsche Gegenwartssprache in einem viel komplexeren als dem rein quantitativen Sinn; nämlich insofern, als die aus ihm gewonnenen grundlegenden Befunde mit gebührender Vorsicht so verallgemeinert werden können, daß sie auch zugleich Aussagen über die deutsche Gegenwartsprosa zulassen. Mehr war nie verlangt, mehr war auch nie behauptet worden.

Anmerkungen

1) Für die Auswahl der einzelnen Texte waren jeweils mehrere verschiedene Gesichtspunkte maßgebend.

2) Im allgemeinen überschätzt man freilich die sprachlichen Folgen politisch-sozialer Grenzziehungen. Vgl. dazu Hugo Mosers illustrative und kritische Schrift "Sprachliche Folgen der politischen Teilung Deutschlands", 1962 = Beiheft zum "Wirkenden Wort", 3; auch "Das Aueler Protokoll". Deutsche Sprache im Spannungsfeld zwischen West und Ost, Düsseldorf (Schwann) 1964.

3) Peter von Polenz, zur Quellenwahl für Dokumentation und Erforschung der deutschen Sprache der Gegenwart, in: Satz und Wort im heutigen Deutsch = Sprache der Gegenwart, Band I, S. 363-378.

4) Wie das von der Bonner Außenstelle erarbeitete Corpus, das der Ermittlung der (vorwiegend lexikalischen) sprachlichen Besonderheiten in beiden Teilen Deutschlands dient.